THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

# Data-driven insights on the dissemination of antibiotic resistance genes

**David Lund**

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology
Göteborg, Sweden 2025

Data-driven insights on the dissemination of antibiotic resistance genes
David Lund
ISBN 978-91-8103-222-2

Acknowledgements, dedications, and similar personal statements in this thesis, reflect the author's own views.

© David Lund, 2025

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr 5680
ISSN 0346-718X

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Typeset with LaTeX
Printed by Chalmers digitaltryck
Göteborg, Sweden 2025

# Data-driven insights on the dissemination of antibiotic resistance genes

**David Lund**

Division of Applied Mathematics and Statistics
Department of Mathematical Sciences
Chalmers University of Technology

## Abstract

Antibiotic resistance is increasing among pathogens, representing a serious threat to public health. Bacteria often become resistant by acquiring mobile antibiotic resistance genes (ARGs), which are disseminated via horizontal gene transfer. To anticipate the emergence of new ARGs and limit their spread, we must increase our knowledge about resistance genes that exist in different environments and about their horizontal dissemination among bacteria. The six papers presented in this thesis aim to provide an extensive characterization of the resistome and an analysis of horizontal ARG dissemination. In Paper I, a previously unseen diversity of genes giving resistance to aminoglycoside antibiotics was identified, including 50 previously unknown mobile ARGs carried by human pathogens. In Paper II, the abundance of ARGs, both well-studied and computationally predicted, was estimated in different microbiomes, revealing a widespread presence of previously unknown ARGs across all analyzed environments. In Paper III, a detailed characterization of the resistomes of the human gut and wastewater microbiomes was performed, highlighting the relationship between ARG prevalence in these microbial communities and potential implications for human health. Papers IV and V present a phylogenetic method to identify horizontal ARG transfer between evolutionarily divergent bacteria, which was used to analyze inter-phyla ARG transfers, and combined with machine learning to quantify the impact of different factors on horizontal ARG dissemination. Finally, in Paper VI, the potential use of machine learning to predict the dissemination of emerging ARGs was evaluated. The resulting models showed promise but need further refinement to inform clinical decision-making. Together, the findings presented in this thesis increase our understanding of how ARGs transfer between bacterial species and communities, highlighting the presence in anthropogenic microbiomes and genetic compatibility as key factors associated with successful ARG dissemination. Moreover, the results demonstrate the utility provided by data-driven methods for improving surveillance and diagnostics of antibiotic resistance.

**Keywords:** antibiotic resistance, horizontal gene transfer, microbiome, hidden Markov model, random forest, phylogenetic analysis

## List of publications

This thesis is based on the work represented by the following papers:

I. **Lund, D.**, Coertze, R. D., Parras-Moltó, M., Berglund, F., Flach, C. F., Johnning, A., Larsson, D. G. J., & Kristiansson, E. (2023). Extensive screening reveals previously undiscovered aminoglycoside resistance genes in human pathogens. *Communications Biology*, 6(1), 812. doi: 10.1038/s42003-023-05174-6

II. Inda-Díaz, J. S., **Lund, D.**, Parras-Moltó, M., Johnning, A., Bengtsson-Palme, J., & Kristiansson, E. (2023). Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes. *Microbiome*, 11(1), 44. doi: 10.1186/s40168-023-01479-0

III. **Lund, D.**, Johnning, A., Holmström, M., Varghaei, L., Inda-Díaz, J. S., Bengtsson-Palme, J., & Kristiansson, E. (2025). Community-promoted antibiotic resistance genes show increased dissemination among pathogens. Preprint: *bioRxiv*. doi: 10.1101/2025.05.12.653433.

IV. Parras-Moltó, M., **Lund, D.**, Ebmeyer, S., Larsson, D. G. J., Johnning, A., & Kristiansson, E. The transfer of antibiotic resistance genes between evolutionarily distant bacteria. Accepted for publication in *mSphere* (2025). Preprint: *bioRxiv* (2024). doi: 10.1101/2024.10.22.619579.

V. **Lund, D.**, Parras-Moltó, M., Inda-Díaz, J. S., Ebmeyer, S., Larsson, D. G. J., Johnning, A., & Kristiansson, E. (2025). Genetic compatibility and ecological connectivity drive the dissemination of antibiotic resistance genes. *Nature Communications*, 16(1), 2595. doi: 10.1038/s41467-025-57825-3.

VI. **Lund, D.**, Axillus, S., Larsson, D. G. J., Johnning, A., & Kristiansson, E. (2025). Can we predict the spread of emerging antibiotic resistance genes? *Manuscript*.

Additional papers not included in this thesis:

VII. **Lund, D.**, Kieffer, N., Parras-Moltó, M., Ebmeyer, S., Berglund, F., Johnning, A., Larsson, D. G. J., & Kristiansson, E. (2022). Large-scale characterization of the macrolide resistome reveals high diversity and several new pathogen-associated genes. *Microbial genomics*, 8(1), 000770. doi: 10.1099/mgen.0.000770

VIII. Nilsson, R. H., Jansson, T., Wurzbacher, C., Anslan, S., Belford, P., Corcoll, N., Dombrowski, A., Ghobad-Nejhad, M., Gustavsson, M., Gómez Martinez, D., Kalsoom Khan, F., Khomich, M., Lennartsdotter, C., **Lund, D.**, Van Der Merwe, B., Mikryukov, V., Peterson, M., Porter, T., Põlme, S., Retter, A., Sanchez-Garcia, M., Svantesson, S., Svedberg, P., Vu, D., Ryberg, M., Abarenkov, K., & Kristiansson, E. (2024). 20 years of bibliometric data illustrates a lack of concordance between journal impact factor and fungal species discovery in systematic mycology. *MycoKeys*, 110, 273. doi: 10.3897/mycokeys.110.136048

## Author contributions

I. Participated in the study design. Retrieved bacterial genomes. Created and optimized hidden Markov models, and used them to predict antibiotic resistance genes (ARGs) in the downloaded genomes. Implemented phylogenetic analysis, phylum enrichment analysis, and genetic context analysis of the predicted ARGs. Analyzed the results and generated all figures and tables. Drafted and edited the manuscript.

II. Implemented genetic context analysis of latent genes from the core-resistome and analyzed the results. Generated Table 2. Discussed the results of the study and edited the manuscript.

III. Participated in the study design. Retrieved bacterial genomes and screened them for ARGs. Generated the ARG reference database. Retrieved and filtered metagenomic data. Estimated the abundance of ARGs in the downloaded metagenomes and analyzed the properties of differently promoted ARGs. Analyzed the results and generated all figures and tables. Drafted and edited the manuscript.

IV. Participated in the study design. Retrieved bacterial genomes. Implemented genetic context analysis of ARGs involved in inter-phylum transfers and analyzed the results. Generated Figures 3, S1, and S20. Analyzed the results of the study and edited the manuscript.

V. Participated in the study design. Retrieved bacterial genomes and screened them for ARGs. Retrieved metagenomic data. Implemented the pipeline for generating observed and randomized horizontal transfers of ARGs from phylogenetic trees. Implemented kmer analysis to measure genetic incompatibility. Estimated the co-occurrence of bacteria in metagenomes. Trained and evaluated the random forest models. Analyzed the results and generated all figures and tables. Drafted and edited the manuscript.

VI. Participated in the study design. Retrieved bacterial genomes and screened them for ARGs. Retrieved metagenomic data. Implemented the pipeline for generating observed and randomized horizontal transfers of ARGs from phylogenetic trees. Implemented kmer analysis to measure genetic incompatibility. Estimated the co-occurrence of bacteria in metagenomes. Analyzed the functional content of bacterial genomes. Trained and evaluated the predictive models. Analyzed the results and generated all figures and tables. Drafted and edited the manuscript.

# Acknowledgements

I am deeply grateful to everyone who has supported me throughout my PhD studies. Without your guidance, trust, company, and understanding, I would have never reached this far. I therefore want to take this opportunity to extend some special thanks:

First, and most importantly, I want to thank my main supervisor, Erik Kristiansson. Working with you over the last five years has been a fantastic experience. Your enthusiasm about even the most minor scientific finding is incredibly motivational, and your endless positivity is contagious, even on the most difficult day. I could not have asked for a better supervisor. Next, I want to thank my co-supervisors Anna Johnning and Joakim Larsson. Thank you, Anna, for all the great feedback and emotional support you have provided, and for always being available to answer questions, listen to me rant about something, or recommend some cool music. Thank you, Joakim, for all of the valuable input, interesting discussions, and encouragement that you have provided. I also want to thank my examiner, Fredrik Westerlund.

The work presented in this thesis includes many important contributions from my collaborators, both local, including Juan Salvador Inda-Díaz, Marcos Parras-Moltó, Sophia Axillus Michaela Holmström, and Laleh Varghaei, and from other departments, including Johan Bengtsson-Palme, Fanny Berglund, Roelof Coertze, Stefan Ebmeyer, and Carl-Fredrik Flach. Thank you all for your hard work and expertise. You have been wonderful to work with.

I have been fortunate to be surrounded by great people with whom to share successes and setbacks alike. I therefore want to thank all current and past members of the Kristiansson research group, including Anna, Mikael, Juan, Marcos, Patrik, Astrid, Martin, Styrbjörn, Sophia, Laleh, Helga, and, of course, Erik, for all of the inspiration and laughs that you have given me. Moreover, I want to thank my other colleagues at the Department of Mathematical Sciences for creating a great work environment and for fun discussions over lunch and fika. Special thanks go to Marija Cvijovic's research group, without whom I am unlikely to have ever ended up at the department.

Finally, I want to thank my friends and family for always supporting me, and my wife, Julia, for your patience and unwavering belief in me. I could not have done it without you.

# Contents

**Abstract** **iii**

**List of publications** **v**

**Acknowledgements** **ix**

**Contents** **xi**

**1 Background** **1**

1.1 Antibiotic resistance . . . . . 1

1.2 Horizontal gene transfer . . . . . 4

1.3 Antibiotic resistance in the environment . . . . . 5

1.4 DNA sequence data . . . . . 7

**2 Aims** **11**

**3 Computational analysis of antibiotic resistance genes** **13**

3.1 Identification of ARGs in sequence data . . . . . 13

3.2 Phylogenetic analysis . . . . . 15

3.3 Detection of horizontal gene transfer from sequence data . . . . 18

3.4 Machine learning in bioinformatics . . . . . 20

**4 Summary of papers** **27**

4.1 Paper I . . . . . 27

4.2 Paper II . . . . . 29

4.3 Paper III . . . . . 31

4.4 Paper IV . . . . . 33

4.5 Paper V . . . . . 36

4.6 Paper VI . . . . . 39

**5 Conclusions** **43**

5.1 Future research . . . . . 45

**Bibliography** **47**

**Papers I–VI**

# 1 Background

The discovery of antibiotics and their subsequent introduction into clinical use during the early 20th century represents one of the most important historical advances in human healthcare. Indeed, these compounds represent one of the cornerstones of modern healthcare, as they have not only provided the means for treatment and prevention of infectious diseases, but they have also enabled the development of many contemporary medical and surgical procedures [1]. The mid-20th century, sometimes referred to as the Golden Age of Antibiotics, saw the discovery of numerous different antibiotics. These mainly encompassed naturally occurring antimicrobials produced by bacteria or fungi, although some compounds were instead manufactured synthetically or semi-synthetically [2]. The interest in developing new antibiotics slowed down significantly after the 1960s, however, with only two new classes of antibiotics having been introduced since the end of the Golden Age [3]. As a result, human society has continued to rely heavily on already established antibiotics for clinical healthcare. Unfortunately, bacterial pathogens are gradually becoming immune to the antimicrobial effects of these compounds, a phenomenon known as antibiotic resistance. This poses an obvious threat to human health, as it has the potential to make common infectious diseases more difficult or even impossible to treat [4].

## 1.1 Antibiotic resistance

Antibiotic resistance refers to the ability of microbes to withstand the effects of antimicrobial compounds at otherwise inhibitory concentrations. Effectively, this means that a given antibiotic loses its potency to treat or prevent infections caused by bacteria resistant to that drug [5]. Resistance can arise as a result of different evolutionary processes, which can generally be divided into three categories: intrinsic, adaptive, and acquired resistance. The first,

intrinsic resistance, refers to an increased tolerance to specific antibiotics as a result of some inherent biological property of a given bacterium. Here, the most prominent example is the outer membrane of Gram-negative bacteria, which is impermeable to many classes of antibiotics, resulting in an intrinsic multidrug-resistant phenotype [6]. The second type of resistance is adaptive resistance, which refers to the ability to modulate existing cellular functions in response to antibiotic pressure to achieve transient antibiotic resistance. This has, for example, been shown to result in high levels of resistance in the pathogen *Pseudomonas aeruginosa*, however, the biological processes underlying the emergence of these phenotypes are not well understood [7, 8]. Indeed, most research has focused on the third type of resistance, acquired resistance, which has greater permanence than adaptive resistance, and under the right conditions can be rapidly disseminated among human pathogens [9].

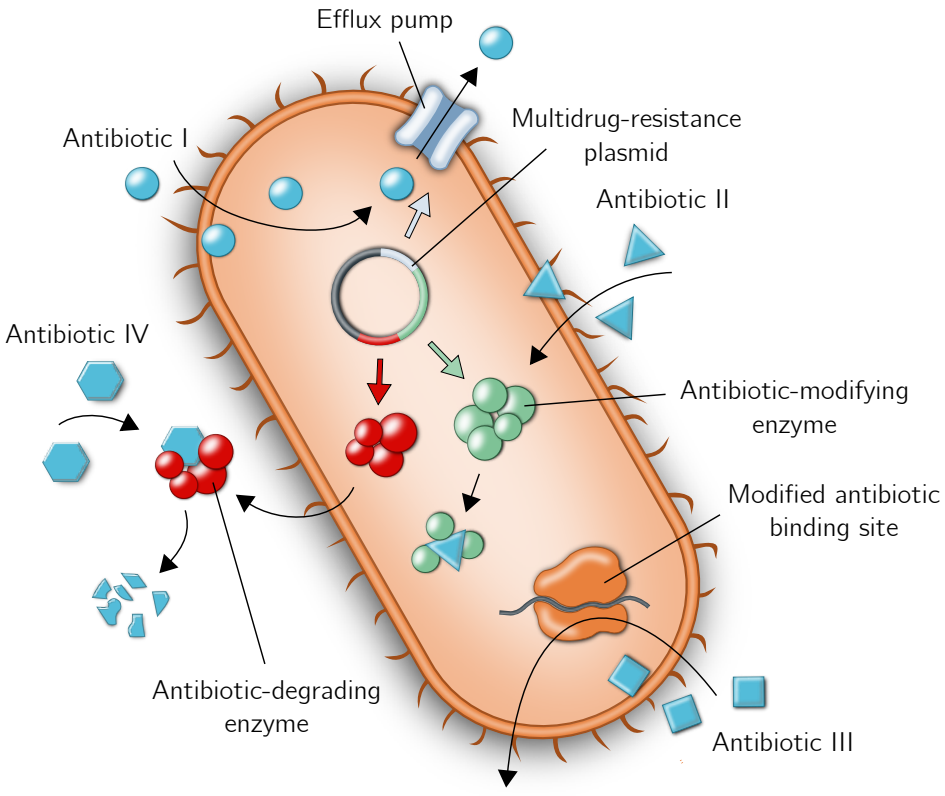



**Figure 1.1:** Illustration of the main mechanisms encoded by antibiotic resistance genes. These include enzymes that break down the antibiotic, enzymes that inactivate the antibiotic by altering their chemical structure, active efflux pumps that transport antibiotics from the cytoplasm to the extracellular matrix, and enzymes that alter the target binding site of the antibiotic.

In general, acquired antibiotic resistance is caused either by mutations in the

bacterial chromosome as a result of adaptive evolution or through the acquisition of antibiotic resistance genes (ARGs). Chromosomal mutations are an important source of resistance in some bacterial pathogens. Here, a prominent example is high-level resistance to fluoroquinolones which most often is a result of mutations in the genes encoding the primary and secondary targets of these drugs [10]. Mutations, however, are limited in their ability to spread among bacteria, since they can only be transferred vertically between parent and daughter cells. By contrast, ARGs can become mobile by transitioning from the chromosomes of their original host bacteria onto mobile genetic elements (MGEs), which are pieces of genetic material that can move independently from the rest of the genome [11]. Once mobile, ARGs can be passed on from the original host to distantly related cells through horizontal gene transfer (HGT). Taking advantage of this, some bacteria have accumulated ARGs from different sources over time into large genetic constructs that, when acquired, confer resistance to many different classes of antibiotics [12].

To date, thousands of ARGs have been identified, each giving increased resilience to anything from a single antimicrobial compound to multiple classes of antibiotics [13, 14]. Consequently, these genes collectively encode a wide range of molecular mechanisms, which are illustrated in Figure 1.1. Broadly, these mechanisms include: (1) reducing the drug's access to the cell, either by decreasing membrane permeability or through active efflux, exemplified by the AcrAB-TolC and Mex multidrug efflux pumps in Gram-negative pathogens [15]; (2) modifying the drug's target, where a functional group or protein occupies the binding site of the antibiotic, such as is the case for Erm 23S rRNA methyltransferases that confer resistance to macrolide, lincosamide, and streptogramin B antibiotics [16]; and (3) directly modifying the antibiotic, where an enzyme chemically inactivates or degrades the drug, exemplified by the beta-lactamase enzymes that hydrolyze beta-lactam antibiotics [17, 18]. Since the introduction of antibiotics, ARGs conferring resistance to all classes of clinically used antibiotics have emerged in human pathogens. Although some of these are intrinsically part of the genomes of certain pathogens, most have been acquired through HGT [14]. In fact, new ARGs are regularly discovered in bacteria causing infections, demonstrating that the mobilization of resistance determinants has not stagnated. However, the evolutionary origins of these genes are still not well understood, which hampers our ability to prevent or delay the establishment of new ARGs among pathogens [19].

## 1.2 Horizontal gene transfer

Most clinically problematic ARGs today are mobile and can thus be shared between bacteria through HGT. This, for example, allows these ARGs to move between harmless commensal species and pathogens, even if they are only distantly related. As a result, HGT is one of the main causes of the spread of resistance within and between bacterial communities, and today the HGT phenomenon is actively studied in relation to antibiotic resistance [20].

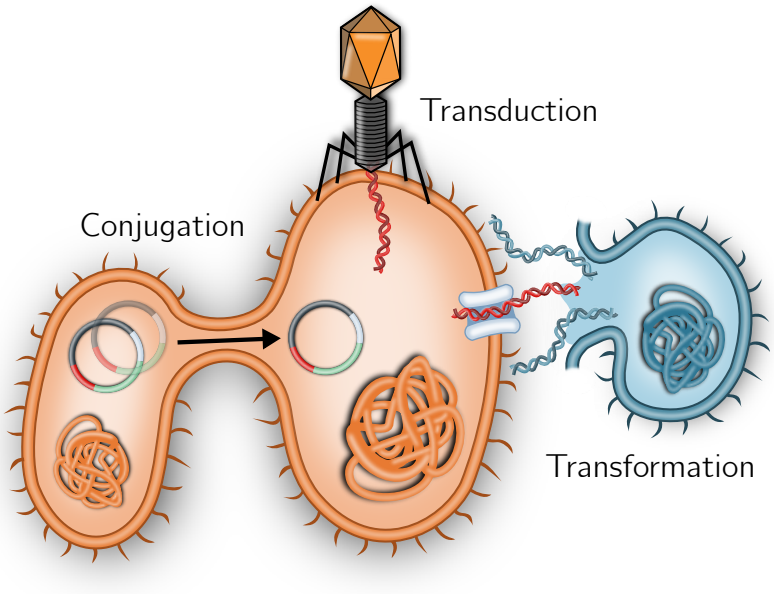



**Figure 1.2:** Illustration of the main horizontal gene transfer mechanisms. Natural transformation, a process where DNA is "released" from one cell and eventually is taken up and incorporated into the genome of another cell. Transduction, a process where bacteriophages are used to transfer genes between two cells. Conjugation, a process where a sex pilus is formed between two adjacent cells, and genetic material is passed from the donor cell to the recipient.

As mentioned above, mobile genes are typically associated with and/or carried by MGEs. There are many different MGEs associated with ARGs, including conjugative elements and insertion sequences/transposons [21]. It is also not uncommon for several MGEs to exist together as part of larger mobile genetic constructs such as plasmids [22]. The movement of DNA between cells has mainly been thought to occur through one of three mechanisms, illustrated in Figure 1.2. These include (1) natural transformation, or uptake of free (non-cell bound) DNA into the cell, (2) transduction, where the transfer is mediated by bacteriophages, and, perhaps most importantly, (3) conjugation, whereby a sex pilus is formed between adjacent bacteria through which the DNA moves

from the donor cell to the recipient [23]. Conjugation requires a specific set of genes to initiate transfer that are often located on conjugative plasmids, together with other genetic material. Where transformation and transduction can occur as side-effects of other biological processes, plasmid conjugation, by comparison, is a more efficient and reliable way for the recipient to acquire foreign DNA directly from the donor [24]. In some instances, conjugation also enables the recipient to develop resistance towards multiple antibiotics through a single HGT event, by acquiring a large multidrug-resistance plasmid [25]. It should also be noted that HGT can also be mediated by other mechanisms [26]. Illustrating this, studies have shown that bacteria are able to transmit genetic material through the use of membrane vesicles [27], nanotubes (tiny pilus-like structures) [28], and phage-like gene transfer agents [29].

Although the main mechanisms associated with HGT have been relatively well studied, much remains unclear about the effect of different factors on the horizontal transfer of ARGs. This includes both genetic factors that would make certain species more likely to engage in HGT, as well as environmental factors, namely in what type of environment(s) horizontal ARG transfer is likely to occur. Indeed, it is well known today that ARGs are ubiquitously present in members of many different microbiomes [30], however, where these ARGs are most likely to mobilize and spread from their original host remains unknown. It is clear that we need to increase our knowledge about the dissemination of ARGs through HGT to combat the spread of new forms of multidrug-resistant pathogens.

## 1.3 Antibiotic resistance in the environment

While the evolutionary details of most ARGs remain unclear, it is known that these genes existed well before humans started treating infections with antibiotics. In fact, some ARGs are suggested to have first evolved billions of years ago [31]. Since they first arose, ARGs have evolved and diversified over long periods of time, which has resulted in the vast resistome, i.e., the complete collection of ARGs carried by microorganisms, which can be observed today [32].

In addition to the resistance determinants commonly encountered in pathogens, the current resistome encompasses a genetic diversity that far exceeds what has been observed in clinical settings to date. These diverse ARGs are found mainly in the genomes of non-pathogenic bacteria that inhabit different environments, including both external environments such as soil and water, and host-associated environments such as human microbiomes [33]. Although the

resistome evolves independently of human interference, the process has been expedited by the excessive use of antibiotics by human society over the last century. The increased concentrations of antibiotics in different environments have provided enough selection pressure for resistance genes to develop, mobilize, and transfer within and between bacterial communities at rates that were likely not reached in the pre-antibiotic era [34, 35]. Furthermore, as these different environments interact, for example, by humans eating animals or by human excrement ending up in the ocean, mobile ARGs are able to flow between different environments, as illustrated in Figure 1.3. Effectively, this means ARGs originating in any environment have the potential to emerge in pathogens and become clinical problems [30].

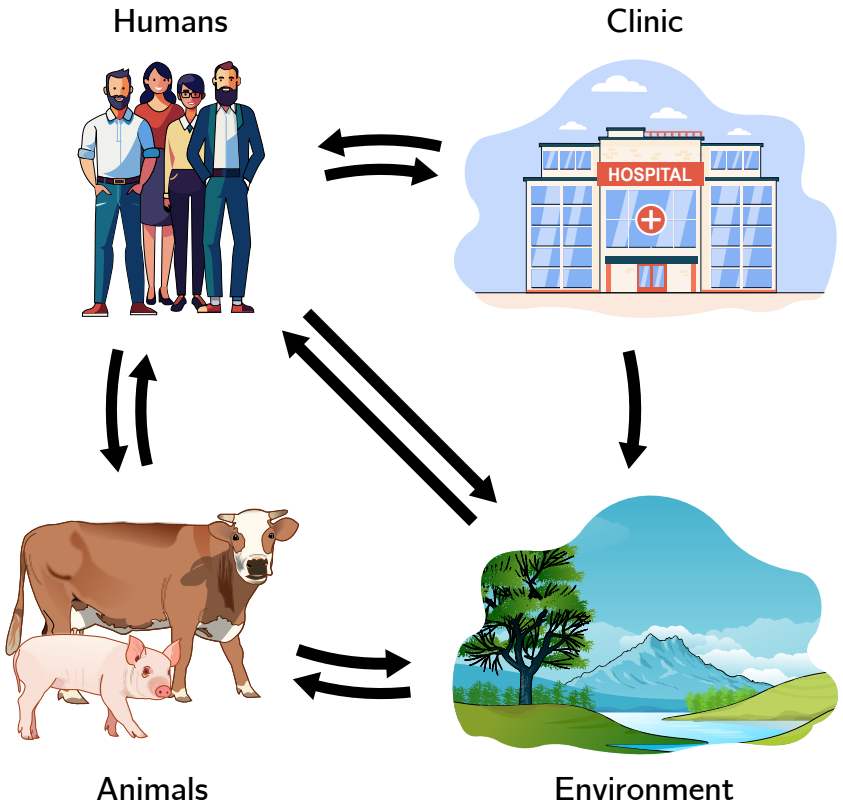



**Figure 1.3:** Illustration of the flow of antibiotic resistance genes between humans, animals, the environment, and the clinic.

To date, many studies have been performed with the aim of characterizing the resistome [36, 37, 38]. In particular, the human gut and sewage microbiomes have been extensively studied with regard to ARGs, due to the known presence of common pathogens and the higher-than-average antibiotic selection pressure associated with these microbiomes [39, 40]. Due to these favorable conditions

for the proliferation of antibiotic resistance, previous studies have described these environments as hotspots where ARGs are likely to emerge in pathogens [33]. However, less emphasis has been placed on analyzing many other types of environments. This has led to bias in the current sequence repositories, where samples collected from external environments are highly underrepresented [41]. To obtain a more complete understanding of how ARGs flow from non-pathogens to pathogens, it is important that the genetic reservoir present in external environments is not overlooked.

## 1.4 DNA sequence data

Genome sequencing is a fundamental part of any contemporary research on microorganisms. The genome of every living organism on Earth is made up of the same four nucleic acids: adenine (A), cytosine (C), guanine (G), and thymine (T), and each organism maintains a copy of its genome in its cell(s). The observable traits of each organism are then determined by the specific sequence of nucleic acids encompassing their genome, where specific regions, or genes, are transcribed into RNA and then translated into proteins [42]. In this context, sequencing refers to the process of identifying the nucleic acid sequence that makes up a given piece of DNA (or RNA) [43].

The first successful results of DNA sequencing were published by Holley et al. in 1965, and analyzed a tRNA molecule isolated from yeast [44]. It would take until 1995 for the first full bacterial genome to be sequenced [45], which was followed by the first draft of the human genome in 2001 [46]. These early advances in whole genome sequencing (WGS) were achieved using the Sanger method, which, although highly accurate, is very costly and has low throughput. Consequently, this method became mostly obsolete for WGS purposes by the advent of next-generation sequencing (NGS) technologies in the 2000s. These new methods greatly increased throughput by generating millions of reads in parallel, and, as a result, the cost per sequencing read decreased dramatically [43]. The generation of NGS data has become even less expensive with time, as technology has been further refined, greatly increasing the availability of DNA sequencing to the research community [47]. Thus, the "third-generation" sequencing methods that debuted in the 2010s were not aimed at further pushing the throughput. Instead, these techniques focused on generating longer reads without the need for pre-amplification of the DNA [48].

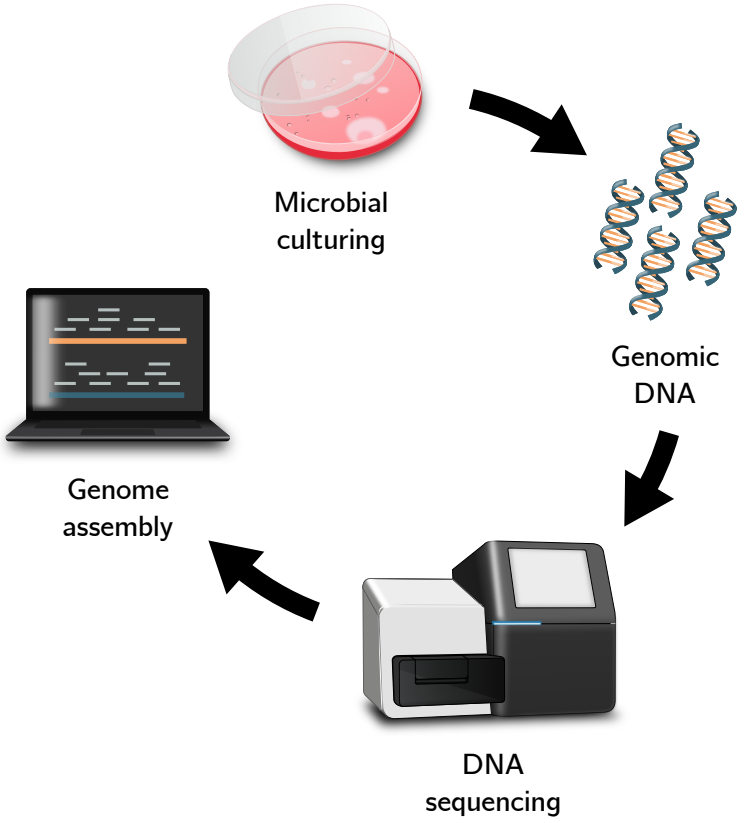



**Figure 1.4:** Flowchart depicting microbial whole genome sequencing. Cells are first cultured in the lab, after which their DNA is extracted, sequenced, and assembled.

Today, vast amounts of WGS data have been generated and deposited in public databases. Exemplifying this, the NCBI Assembly database of draft genomes currently contains more than 2.5 million sequenced genomes from bacteria alone [49]. By mining such data, we have been able to study biological phenomena in a way that was not possible before the advent of DNA sequencing. Indeed, many data-driven methods have been developed in recent years for studying antibiotic resistance, which has significantly advanced our understanding of the topic [50].

### 1.4.1 Metagenomics

Since the inception of microbiology, we have relied primarily on culture-based methods for the characterization of microbes. As the name suggests, these methods require that the microorganism(s) of interest be cultivated in the laboratory before they can be studied [51]. This is also true for WGS, as illustrated in Figure 1.4. Although these methods have allowed for the characterization of

e.g. thousands of bacterial species, the vast majority of microbes are unfortunately not able to be cultured under standard laboratory conditions, and so they remain elusive. In fact, to date, it has been estimated that we have discovered no more than 1% of the total microbial diversity on Earth [52]. To mitigate this problem and improve our understanding of the unknown microbiome, several molecular methods have been developed that do not rely on culturing [53]. Among these methods, metagenomic sequencing is arguably the most notable.

In contrast to more conventional DNA sequencing methods that rely on isolating the DNA of a specific organism before sequencing, metagenomic sequencing instead applies a more brute-force approach, sequencing all DNA present in a given sample [54]. Generally, metagenomic sequencing is used for one of two purposes: 1) to analyze the taxonomic composition of microbial communities and 2) to estimate the abundance of different genes in microbial communities. The first of these is most often achieved through a method called *amplicon sequencing*. This method is based on the identification of specific marker genes that are present in all microbes of a specific type (e.g., bacteria). For a gene to serve as an effective marker, it must have highly conserved regions for primer design (to enable PCR amplification before sequencing) and highly variable regions for taxonomic identification. The most widely used marker for bacteria is the 16S rRNA gene, which meets both criteria [55]. After sequencing, the different 16S rRNA genes present in a sample can be clustered at 97% identity into what are known as operational taxonomic units (OTUs). In essence, each OTU represents a putative species, which can be assigned a taxonomic affiliation based on reference sequences from databases. It has, however, been suggested that the default 97% cut-off results in a merging of species, which still leads to an underestimation of microbial diversity [56]. As an alternative to OTUs, methods have been developed to infer amplicon sequence variants (ASVs) in a sample that can differ as little as one nucleotide [57]. Conversely, this approach has been suggested to have the opposite problem, where the increased resolution results in the separation of a single genome into multiple ASVs [58]. Nevertheless, amplicon sequencing has been shown to provide a considerably more accurate view of microbial populations compared to culture-based methods [59].

The other type of metagenomic sequencing is called *shotgun sequencing*. In contrast to amplicon sequencing, shotgun sequencing is completely non-targeted and works by sequencing randomly from the pool of genetic information in a microbial sample using high-throughput sequencing. The resulting reads can then be used to assemble metagenomic contigs or mapped directly to a set of reference genes, providing information about the genetic content present in a given microbiome [54]. Since the popularization of metagenomics, several massive initiatives have been carried out with the aim of characterizing the

unknown microbiomes throughout the world. Examples include the Human Microbiome Project [60] and the Tara Oceans project [61], which collectively have produced over 100 TB of metagenomic sequencing data. Metagenomics has also been widely used to study the resistome, the vast reservoir of antibiotic resistance genes carried by bacteria [62, 63, 64], and this method has provided new insights into the environmental dynamics of antibiotic resistance.

Together, the vast amounts of WGS and metagenomic data available from repositories like NCBI [49] and MGNify [65] constitute a remarkable resource. By leveraging this resource to develop data-driven methods for studying biological phenomena, we have an unprecedented opportunity to increase our insight into issues like antibiotic resistance. This could help shape our perspective and develop solutions to the problems that human health is facing.

# 2 Aims

This thesis aims to extend our knowledge about the contents of the unknown resistome, as well as deepen our understanding of the causes behind the successful horizontal dissemination of antibiotic resistance genes (ARGs). This information may prove vital for anticipating the emergence of new ARGs and preventing their uncontrolled spread among pathogens. The six papers that make up this work can broadly be divided into two parts, each of which contributes via the following aims:

1. Characterization of the unknown resistome
    a Identify new mobile resistance genes in pathogens and experimentally validate their functionality (Paper **I**).
    b Explore the resistome in different environments, including both well-known (established) and computationally predicted (latent) resistance genes (Papers **II–III**).
    c Perform a specific investigation of the resistomes associated with the human gut and wastewater microbiomes to identify the environment(s) where clinically relevant ARGs are promoted (Paper **III**).
2. Identification of patterns underlying horizontal ARG transfer
    a Develop and apply a method to investigate the propensity of ARGs to transfer between evolutionarily divergent bacteria (Papers **IV–V**).
    b Estimate the influence of different genetic and environmental factors on the horizontal transfer of ARGs (Paper **V**).
    c Investigate the potential of machine learning for predicting the future dissemination of novel ARGs (Paper **VI**).

# 3 Computational analysis of antibiotic resistance genes

This chapter provides a brief description of the main methods that were used in the papers presented in this thesis.

## 3.1 Identification of ARGs in sequence data

A large number of antibiotic resistance genes (ARGs), conferring resistance to all classes of antibiotics used for clinical infection treatment, is known to circulate among human pathogens [66]. When new resistance genes emerge, they are typically discovered after causing resistance in a clinical isolate, at which point they likely already have spread widely among bacterial communities. For example, this was the case for the beta-lactamase NDM-1 and the colistin resistance determinant MCR-1 [67, 68]. To overcome the drawbacks of traditional surveillance, several computational methods have recently been developed that can identify ARGs, including new variants, from whole genome sequencing (WGS) and metagenomic sequencing data [50].

### 3.1.1 Identification of ARGs in whole genome assemblies

As new resistance genes have been discovered, their sequences have been collected in databases such CARD [14] and ResFinder [13], which collectively contain thousands of reference ARG sequences at the time of writing. The methods used for the identification of ARGs in sequencing data are generally based on these reference sequences. A widely used approach involves alignment-based homology searches against one or more reference ARG databases using bioin-

formatic tools such as BLAST or bowtie [69]. This relatively simple approach is useful for annotating genes that are very similar to the known reference ARGs, however, it fails to identify homologs that are less evolutionarily close to the genes in the databases. Since much of the resistome consists of such homologs, this means that many potentially problematic ARGs are overlooked [70].

To address this problem, more sophisticated methods have been developed, which apply different computational frameworks for the prediction of resistance genes. In general, these methods use the reference ARG databases to create models that can also identify more distantly related homologs to the reference genes based on similarities in gene sequence or protein structure [71]. One of the most well-established computational frameworks for gene prediction is based on profile hidden Markov models (HMMs). These models are built from multiple sequence alignments and can identify homologous genes based on conserved genetic regions rather than overall sequence similarity [72]. This allows for the discovery of previously unknown gene variants but is restricted to the identification of genes that share an evolutionary history with the reference genes used to build the models. One method that uses profile HMMs is fARGene, which in addition to the prediction of ARGs in WGS data also enables gene prediction in metagenomic data without the need for prior assembly [70]. This method has repeatedly shown a high performance for predicting functional new ARGs, in addition to their well-characterized counterparts, from a variety of datasets [73, 74, 75], proving the reliability of HMM-based ARG predictions.

More recently, methods have been developed that apply machine learning algorithms for predicting ARGs. A prominent example is deepARG, which takes a deep learning approach, and uses algorithms and models that can discriminate between true ARGs and genes that contain some ARG-like regions without conferring resistance [76]. Another example is PCM, a method that uses machine learning to make predictions based on protein structure [77]. Similar to the HMM-based methods, the machine learning models are also able to identify previously uncharacterized homologs but are unable to predict ARGs associated with novel resistance mechanisms.

### 3.1.2 Identification of ARGs in metagenomes

In contrast to whole genome assemblies, shotgun metagenomic datasets consist of fragmented DNA originating from many different organisms, which makes the identification of specific genes more complicated [78]. To circumvent this, it is possible to first assemble longer sequences, or even complete genomes, from the metagenomic reads through a process called *de novo assembly*, however,

this is computationally expensive and may become unfeasible when the data grows large [79]. Furthermore, de novo assembly of mobile genetic elements, where many clinically relevant ARGs are located, is highly difficult due to the many repetitive regions that they typically contain [80]. Therefore, several alignment-based methods have been developed that can be applied directly to the raw metagenomic reads.

Arguably the simplest approach for identifying ARGs in metagenomes involves aligning the reads against a curated ARG reference database such as the aforementioned CARD [14] and ResFinder [13], using bioinformatic tools like BLAST [81] or DIAMOND [82]. Here, strict alignment criteria must be used to ensure a low false positive rate; however, this also comes at the cost of a high false negative rate [83]. To amend this, more sophisticated methods specifically designed to identify ARGs in metagenomes have been developed, such as ARGs-OAP [84] and GROOT [85], which are limited to the identification of ARGs already present in reference databases, as well as deepARG [76] and fARGene [70], which can also identify novel ARGs. Here, fARGene is particularly notable as it is also able to reconstruct full-length genes directly from metagenomic reads [70]. More recently, methods that do not rely on sequence alignment have been developed, such as ARGNet, which applies deep neural networks trained on reference sequences to classify ARGs in metagenomes [83].

## 3.2 Phylogenetic analysis

When new genes are discovered, one of the main ways by which we are able to understand them is by studying their evolutionary relationships with other, similar genes. As we cannot directly observe their evolutionary history, we instead infer it through computational phylogenetics. The aim of phylogenetic analysis is to divulge the evolutionary relationships of genes or taxa through the reconstruction of phylogenetic trees, representations of the evolutionary tree computed from molecular sequences. Over time, a variety of increasingly efficient and sophisticated methods have been developed to meet the demands imposed by the increasing data volumes [86].

A phylogenetic tree is in essence a branching diagram that illustrates the evolutionary relationships between biological sequences. When discussing these diagrams, the observed sequences from which the tree was built are termed *leaves*, positioned at the tips of the structure. The leaves are attached to *branches*, which in turn are connected by *nodes*. Each node represents an inferred common ancestor, and a *clade* refers to all leaves descending from a given node.

The oldest of the nodes, from which all others descend, is denoted the *root*, and can be inferred by the tree-building algorithm or deliberately placed based on prior evolutionary assumptions [87] (Figure 3.1). The root provides the tree with an evolutionary direction, but is not required for constructing a tree. Unrooted trees, however, do not provide information about the evolutionary trajectory of the leaves, only their relatedness [88].

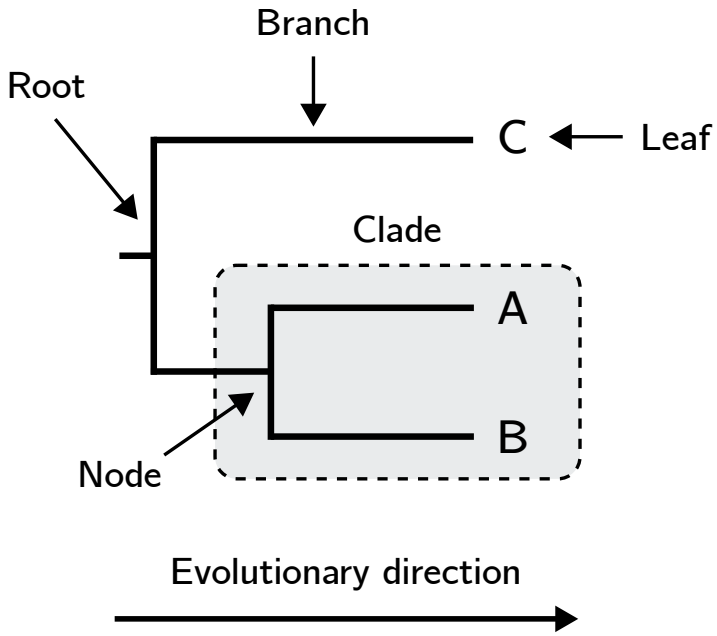



**Figure 3.1:** A basic illustration of a phylogenetic tree, highlighting the standard nomenclature used to describe its different components.

Phylogenetic tree reconstruction methods can broadly be classified into two categories: distance-based methods, such as neighbor-joining and least-squares, and character-based methods, including maximum parsimony, maximum likelihood, and Bayesian algorithms. The distance-based methods, as the name suggests, derive phylogenetic trees from the genetic distances between sequences, which are computed from a multiple sequence alignment. Under the assumption that the sequences that can be observed today accurately reflect every historical genetic divergence event, the true evolutionary tree can be reconstructed from the distances (or amount of dissimilarity between two aligned sequences). The most popular distance-based phylogenetic analysis method is the neighbor-joining algorithm [89].

Briefly, neighbor-joining works by first assuming an un-rooted, bifurcating tree with $N$ leaves. A pair of neighbors consists of two leaves connected via a single interior node. The topology of the tree is determined by iteratively merging the most similar neighbor pairs, forming new pairs at each iteration until a consen-

sus topology is achieved. At each iteration, a distance matrix $D$ is computed from all pairwise distances between the leaves. For each pair of leaves $a, b$, the matrix is then used to calculate the sum of branch lengths resulting from their merging, based on least-squared estimates. The pair producing the smallest sum are merged into a combined leaf $(a, b)$. The distance between $(a - b)$ and another leaf $c$ is given by

$$D_{(a-b)c} = \frac{1}{2}(D_{ac} + D_{bc}) \tag{3.1}$$

which is used to generate a new distance matrix. At each iteration the number of leaves $N$ is reduced by 1, and the process is repeated until $N = 3$, when only a single un-rooted tree topology remains [90].

While distance-based methods are simple and computationally efficient, the assumptions they are based on are questionable at best. Indeed, genetic mutations can be reversed over time, at which point the historic divergence is no longer observable. Consequently, these methods generally perform worse than the more complex character-based methods such as maximum likelihood (ML)-based and Bayesian methods, which derive the tree topology from probability-based algorithms based on a predefined model of sequence evolution [91].

The ML-based methods aim to identify the topology that makes the observed data most probable under a given substitution model. This involves two main steps: first, the likelihood $L(\theta)$, where $\theta$ is an unknown parameter relating to the substitution model parameters and branch lengths, is maximized for each possible tree topology. Next, the topology producing the highest likelihood — i.e., the tree that best explains the observed data — is identified and selected as the most "correct" representation of the evolutionary history. Note, however, that this topology is not guaranteed to accurately reflect the true evolutionary trajectory of the sequences in question [89, 92]. If, for example, the substitution model is poorly chosen, the performance of ML-based models will be negatively impacted, which, in turn, can lead to wrongful conclusions [93]. Moreover, most ML-based methods assume that mutations at each site and lineage happen independently, and, consequently, that the likelihood is the product of the probabilities observed at each site. This assumption, again, does not necessarily align with the biological reality [94, 95].

Phylogenetic methods based on Bayesian statistics are closely related to ML-based approaches. The Bayesian framework is based around the posterior probability P(T|D), which represents the probability that a given tree topology $T$ is correct given the observed data $D$, the prior probability P(T), and a likelihood

function P(D|T), where

$$P(T|D) = \frac{P(T)P(D|T)}{P(D)}. \tag{3.2}$$

Although the posterior probability is simple to define, it is generally not feasible to calculate analytically due to the high dimensionality of the tree topologies and model parameters. As a result, algorithms such as Markov chain Monte Carlo are usually applied to approximate the posterior instead [96, 97].

Like in many other areas of computational biology, additional methods for phylogenetic analysis based on machine learning have also been developed in recent years. These methods have proved to be highly flexible, however, the advantages they provide over more traditional methods have yet to be fully shown [98].

## 3.3 Detection of horizontal gene transfer from sequence data

As discussed in Chapter 1, horizontal gene transfer (HGT) is central to the spread of antibiotic resistance. Consequently, when studying new ARGs it is of high interest to identify those that have undergone HGT, especially if they have transferred into pathogens where they pose a more immediate threat to human health [99]. The most widely used strategy for detecting mobile genes involves searching the genetic regions flanking the gene in question for mobile genetic elements (MGEs) [39]. However, several computational methods have also been developed to identify HGT directly from sequence data. These methods can generally be divided into two categories: parametric and phylogenetic methods [100].

Parametric methods for identifying HGT aim to find genetic regions that exhibit significant deviations from the average characteristics of the host genome. Such deviations suggest that the region originates outside of the genome and, thus, has been acquired through HGT. Here, commonly analyzed characteristics include nucleotide composition, codon bias, and structural features [101]. These methods rely on the assumption that the genome of a given species has been shaped by specific evolutionary pressures, resulting in each species having developed a recognizable genomic signature. Genes that have been acquired from different organisms are thus unlikely to conform to the structure of their

new host genome [102]. The nucleotide composition of a genome is often represented by its GC-content — the proportion of genomic DNA made up of either guanine or cytosine — which can range between less than 20% to over 70% in bacteria [103]. However, while the GC-content can differ substantially even among closely related species, it can also be very similar among distantly related ones. As a result, this estimate has a relatively low resolution, and by relying solely on GC-content some instances of HGT will not be detectable. A more refined approach is based on analysis of codon frequencies, as genomes with similar nucleotide composition can exhibit differences in their preferred codon usage [104]. Compared to genomic GC-content, the codon bias is more difficult to compute, but it can be modeled using methods such as Markov chains [105].

The alternative approach for inferring HGT events is the phylogenetic approach. As the name implies, these methods rely on phylogenetic trees to detect discrepancies that are not explainable by vertical evolution. Specifically, when a gene has undergone HGT, the phylogenetic gene tree (describing the gene's evolutionary history), will conflict with the species tree (representing the evolution of the host organisms). Instead, the transferred gene will be displaced relative to other genes from its host species (Figure 3.2). Because constructing large phylogenetic trees from complete genomes is generally not feasible, species trees are usually derived from well-conserved housekeeping or informational genes [106].

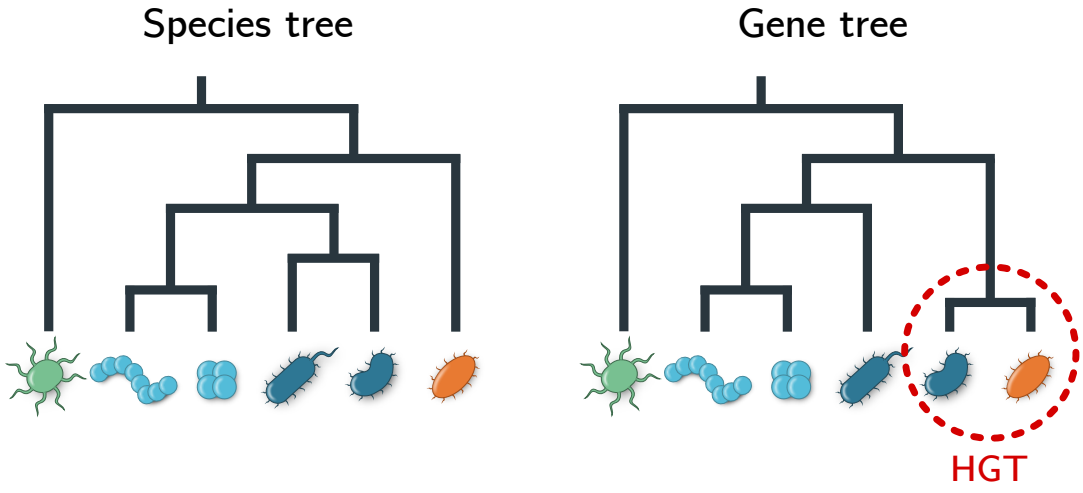



**Figure 3.2:** Illustration of a basic gene tree and corresponding species tree. Encircled is an inconsistency between the two phylogenetic trees, which cannot be explained by vertical evolution and is thus inferred as horizontal gene transfer (HGT).

Phylogenetic methods for detecting HGT can be further classified into two

subcategories: explicit and implicit methods [100]. Explicit methods directly compare gene trees and species trees, for example by applying statistical tests at every site in the trees to identify significant disagreements [107]. However, these methods can quickly become computationally demanding when the trees grow larger. To circumvent this limitation, implicit phylogenetic methods can be applied instead. Although based on similar principles as their explicit counterparts, implicit methods do not rely on a fully constructed species tree. The hosts' evolutionary relationships are instead inferred from sequence similarity or measures of the evolutionary distance of the host species [108]. While HGT between evolutionarily divergent species can be inferred from taxonomy, transfers between more closely related organisms may be more difficult to detect [106]. In such cases, a more sensitive estimate of evolutionary distance can be calculated from a pairwise sequence alignment using Maximum-Likelihood. This estimate can then be evaluated using a statistical test (likelihood-ratio test) to determine whether the observed divergence is significant enough to suggest horizontal transfer [109].

## 3.4 Machine learning in bioinformatics

The vast expansion of biological sequence data over the past two decades has created an unparalleled opportunity to increase our knowledge of biological systems [47, 65, 110]. In order to take full advantage of this data, however, there has been an ever-increasing need to develop new methods for analyzing it. It is no wonder, then, that machine learning has been applied in many areas within the field of bioinformatics to solve problems that would be difficult, if not impossible, to solve using more traditional methods [111].

Machine learning is a broad term used to describe various algorithms that can learn from data. Models created using these algorithms can be used to identify patterns in the data that are too complex for the human eye to detect, and, if trained correctly, apply what they have learned to make predictions about new data [112]. To date, many machine learning algorithms have been developed, ranging from simple, such as linear regression models, to highly complex, such as the recently introduced large language models (LLMs), which can encompass hundreds of billions of parameters [113, 114]. These algorithms can generally be divided into two different categories based on the task they are meant to solve: supervised learning, in which the model is trained on labeled data, with the task of assigning a label (response variable) given a set of other variables, and unsupervised learning, where models are trained to identify patterns in unlabeled data [115].

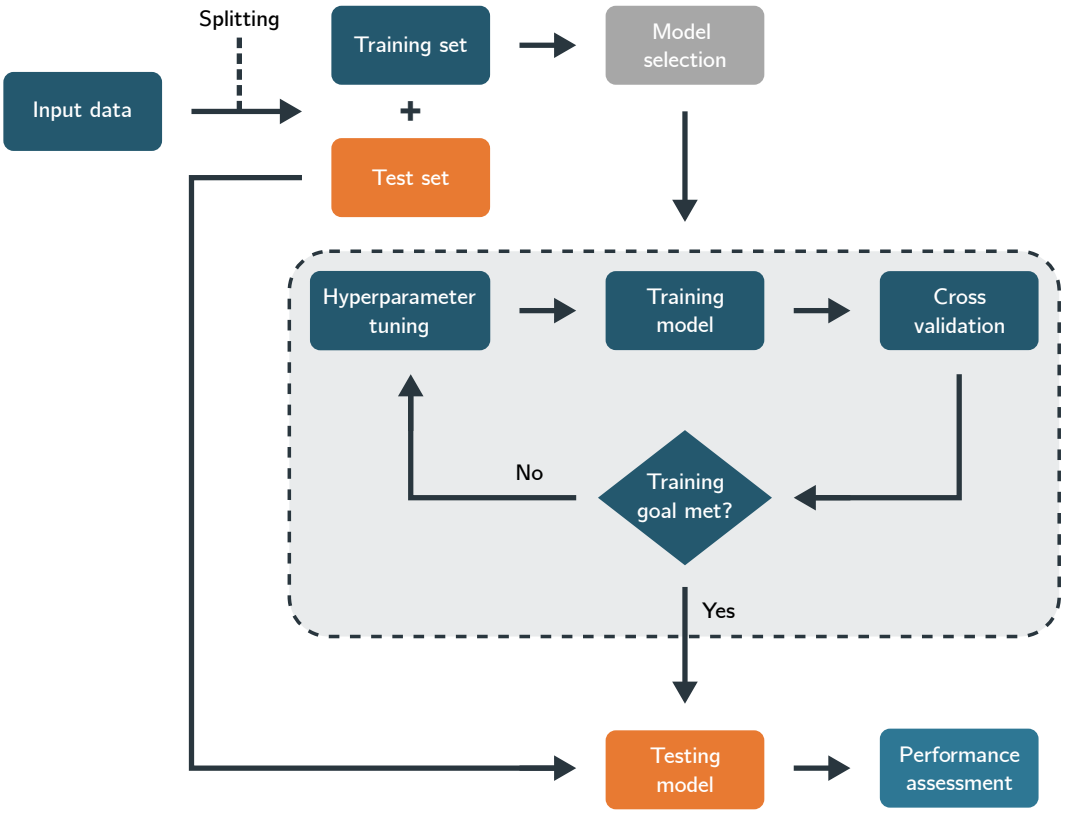



**Figure 3.3:** Workflow for training supervised learning models. The total input data is split into training and test datasets. Training and cross-validation are performed using the labeled training set, while the unlabeled test set is used to evaluate the final model.

Supervised learning is the most widely used form of machine learning, which can be further divided into classification (categorical response variable) and regression models [113]. The general approach to training supervised models is shown in Figure 3.3. Briefly, the total input data is split into a training set and a test set. During the training phase, which generally also includes cross-validation, the model is only learning the features of the training set, including its response variable. After the training has concluded the model is evaluated based on its ability to correctly assign response variables to observations from the test set, from which the labels have been removed. Here, the evaluation of machine learning models can be based on several metrics depending on the type of model. For classification models, popular options include accuracy (proportion of correct predictions), precision (proportion of true positive predictions), sensitivity/recall (proportion of correctly predicted positive observations), and specificity (proportion of correctly predicted negative observations) [116]. Conversely, popular metrics used to evaluate regression

models include the $R^2$ (proportion of variance in the response variable that can be predicted by the predictor variables), as well as the mean absolute error (MAE), the mean squared error (MSE), and the root mean squared error (RMSE), all of which are based on the average distance between predicted and actual values [117]. A properly trained model should be able to accurately label the unseen test data. However, occasionally the training will result in the model overfitting, i.e. learning too much of the noise and random fluctuations present in the training data which does not generalize to other observations [118].

As mentioned earlier in this chapter, machine learning has found applications within genomics and phylogenetics; however, its general usefulness has expanded to other areas of bioinformatics, including proteomics, transcriptomics, metabolomics, and systems biology [111, 119]. Within these areas, machine learning has been used for a large number of tasks, including classification, clustering, prediction, identification of associations, groups, and deviations, and visualization [120]. Illustrating this, machine learning has been successfully used to increase our understanding of host-microorganism interactions for infectious disease research and drug discovery, identify correlations between gut microbiome composition and colorectal cancer, and predict antibiotic resistance phenotypes in pathogens [112]. Arguably, however, the most significant application of machine learning in a biological context to date has occurred within the field of protein structure prediction, which was completely revolutionized by the unveiling of AlphaFold2 in 2020 [121]. The AlphaFold2 algorithm takes a deep learning approach to predict the three-dimensional structure of proteins from their amino acid sequence, predicting protein structures with almost experimental precision and far outperforming any competing software [122]. This biological application of machine learning was considered so significant that it was awarded the 2024 Nobel Prize in Chemistry [123].

### 3.4.1 Random forest models

Among machine learning frameworks used for bioinformatic applications, random forest (RF) models have become one of the more popular choices due to their high accuracy, interpretability, and ability to handle complex, high-dimensional data [124]. As the name suggests, RFs are based on an ensemble (or "forest") of independent decision trees [125]. In these decision trees, each individual node represents a logical test (called a split), and each leaf represents a prediction (Figure 3.4a). For each observation, the outcome is determined by traversing the tree from the root to the leaves, along a path that is determined by its features [126]. Decision tree models have several desirable properties,

including the fact that they are highly intuitive, however, their discrete nature implies a high degree of prediction variance. This lack of robustness can be overcome by combining $M$ independent decision trees into an ensemble and averaging their predictions [127] (Figure 3.4b).

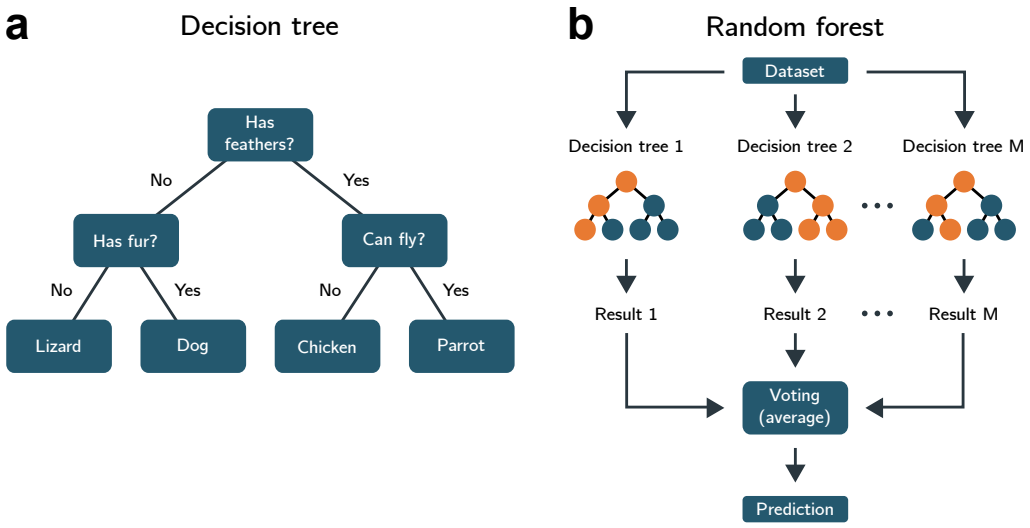



**Figure 3.4:** Illustration displaying the structure of tree-based machine learning models. **a** Example of a basic decision tree. **b** Overview of the random forest architecture.

Originally introduced by Leo Breiman over 25 years ago, the RF algorithm is a versatile ensemble method capable of handling both regression and classification tasks. Its name reflects the incorporation of randomness in the construction of decision trees—specifically, through the random selection of feature and sample subsets used to build each tree [128]. Given a random vector $X = (X_1, ..., X_N)^T$ representing the input variables, or features, and a random variable $Y$ representing the response, we assume an unknown joint distribution $P_{XY}(X, Y)$. The training goal is to find a function $f(X)$ that predicts $Y$, by minimizing the expected loss $E_{XY}(L(Y, f(X)))$, where $L$ is a suitable loss function. For ensemble models, the prediction function $f(x)$ is formed by combining a collection of base-learners $h_1(x), ..., h_M(x)$. How these are combined depends on the task, with regression using an average of the base learners

$$f(x) = \frac{1}{M} \sum_{i=1}^{M} h_i(x) \tag{3.3}$$

while classification uses the consensus vote, i.e. the most frequently predicted

class

$$f(x) = \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{M} I(y = h_i(x)). \tag{3.4}$$

For RFs, each base learner $h_i(X, \Theta_i)$, $i = 1, ..., M$ is an individual decision tree, where $\Theta_i$ represents an independent collection of features [129]. The number of trees making up the ensemble, *M*, is, thus, one of the main hyperparameters that should be considered when training an RF model, alongside the general structure of the decision trees (i.e. how many variables are tested at each node and the minimum number of observations required for each leaf) [130]. Importantly, RF models generally do not need to be combined with additional cross-validation to optimize these hyperparameters, as an estimate of the prediction error — called the "Out-Of-Bag" (OOB) error — is generated as part of the training process. Briefly, each tree is generated using a random bootstrap sample containing $\sim 63\%$ of the data. After the model training is complete, the OOB error is computed by averaging predictions made for each observation using only the trees where that observation was not included in the bootstrap sample. This estimate serves the same function as cross-validation, as it provides an independent assessment of the performance, and can for example be used to optimize the model parameters [125].

In bioinformatics, machine learning is often used not only for prediction but also to gain insight into biological processes. Therefore, the interpretability of the models used is often of high importance [124]. Here, random forests are highly useful since they offer internal estimates of feature importance which can be used to identify the most predictive feature(s) [128]. Various metrics can be used for estimating the feature importance, but the default choice in many software implementations is the so-called Gini importance [131]. When the trees in an RF model are grown, the splitting, or variable(s) tested at each node, is designed to maximize the decrease in impurity introduced by each split. For classification, the impurity is typically measured as the Gini impurity:

$$\hat{\Gamma}(t) = \sum_{k=1}^{K} \hat{\phi}_k(t)(1 - \hat{\phi}_k(t)), \tag{3.5}$$

where $\hat{\phi}_k(t)$ is the class frequency for class $k$ in node $t$ [132]. The Gini impor-

tance of a predictor variable $X_j$ is calculated as the average decrease in impurity at each split formed by $X_j$ across the forest [133]. Another widely used metric is the permutation accuracy importance. This works by randomly permuting the values of a predictor variable $X_j$, thereby disrupting its relationship with the response variable $Y$. The model is then tasked with making predictions using the permuted $X_j$ alongside the other unaltered predictors. The importance of $X_j$ is quantified as the decrease in prediction accuracy resulting from the permutation, where a larger decrease suggests a stronger association between $X_j$ and $Y$ [134].

Identifying the most important predictor variables is key to interpreting machine learning models such as RFs. However, feature importance does not provide information about the marginal effects of these important variables on the predictions made by the model. Instead, other methods are required to extract this information, with arguably the most popular alternative being analysis of partial dependence (PD) [135]. This is a model-agnostic framework that is generally used to produce low-dimensional visualizations of the prediction function, so-called PD plots, which explain the relationship between the model outcome and predictors of interest [136]. Briefly, the method works by fixing the value of a predictor variable $x_j$ at some value $v$. The model is then tasked with making predictions using $x_j = v$, with all other variables retaining their original values from the data, and the average prediction over all samples is calculated as

$$\hat{f}(x_j = v) = \frac{1}{n} \sum_{i=1}^{n} f(x_j = v, x_{i,-j}), \tag{3.6}$$

where $x_{i,-j}$ represents all variables except $x_j$ for sample $i$. This is then repeated multiple times using different values for $v$, and the results are typically visualized as a plot showing the resulting $\hat{f}$ against $v$ [137]. By combining feature importance with PD analysis, we can create highly interpretable machine learning models and, consequently, leverage such models to gain deeper insight into biological processes.

# 4 Summary of papers

This chapter summarizes the aims and findings of the six papers included in this thesis.

## 4.1 Paper I

Antibiotic resistance genes (ARGs) conferring resistance to every major class of antibiotics used to treat infections have been detected in bacterial pathogens. This issue is exacerbated by the fact that new ARGs keep moving into the clinic from external sources. Often, these are acquired through horizontal gene transfer (HGT) from harmless commensal or environmental bacteria, which are known to maintain a large diversity of ARGs [32]. Currently, the lack of knowledge about the resistome, the complete collection of ARGs carried by bacteria, makes it difficult to anticipate and manage new clinical ARGs. Indeed, new ARGs are usually discovered only after they have become disseminated among pathogens [138], at which point further spread is difficult to prevent. Paper I, *Extensive screening reveals previously undiscovered aminoglycoside resistance genes in human pathogens*, therefore, aimed to expand the knowledge about the aminoglycoside resistome, and to demonstrate how large-scale computational screening can be used for early detection of new ARGs acquired by pathogens through HGT before they spread widely.

To do this, we first created and optimized nine profile hidden Markov models (HMMs) — denoted model A to I — for identification of *aac* and *aph* genes, encoding aminoglycoside-modifying enzymes (AMEs). These models were then used to screen $\sim 1$ million public bacterial genomes for these resistance genes using the software fARGene [70]. In total, this yielded 1,071,815 genes encoding 34,053 unique AMEs, divided into 7,612 AME families (<70% between-family amino acid identity), a diversity of AMEs several times larger than previously

described. Next, we evaluated the mobility of predicted genes on a large scale to identify new AMEs that show evidence of being emerging in clinical pathogens. We retrieved the genetic regions directly up and downstream of all predicted ARGs and screened them for mobile genetic elements (MGEs), including genes involved in plasmid conjugation, insertion sequences (ISs), integrons, and co-localized mobile ARGs. This analysis revealed a total of 50 previously unknown AME families carried by pathogenic host species that were also found to co-localize with MGEs (Figure 4.1).

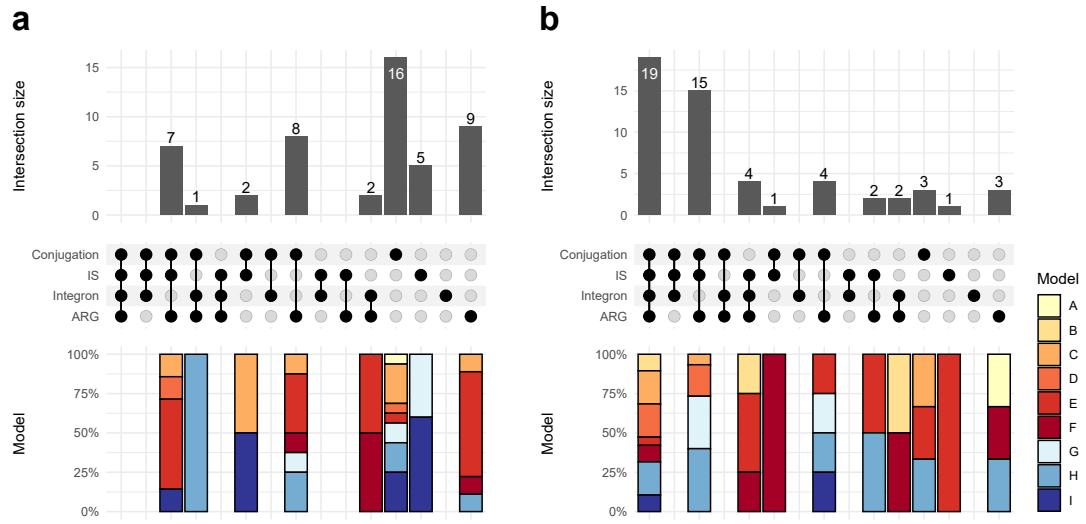



**Figure 4.1:** The number of AME families carried by pathogenic species that were associated with different combinations of genes relating to mobile genetic elements (MGEs), including conjugation systems, insertion sequences (IS), integrons, and/or other known mobile antibiotic resistance genes (ARGs). The bars at the bottom indicate the distribution of genes predicted by each of the nine models within each category. **a** Families representing new AMEs and **b** families representing known AMEs. From Lund et al. 2023 (*Communications Biology*. 2023 Aug 3;6(1):812). Licensed under CC-BY-4.0.

Moreover, genes from 21 of these 50 families were associated with clinical isolates, showing that they have been able to move into pathogens undetected. To confirm the functionality of our predicted ARGs, we selected 28 of the genes associated with both pathogens and MGEs and expressed them in *Escherichia coli*. When the resulting phenotypes were assessed through disk diffusion tests, 21 (86%) of the tested genes produced a resistant phenotype — of which 17 (61%) conferred resistance above the clinical breakpoint(s) and/or epidemiological cut-off value(s) from EUCAST [139] — showing that our models were able to accurately identify previously uncharacterized, potent resistance genes. The results from this paper provide new insights into the aminoglycoside resistome, and demonstrate the usefulness of computational screening as a tool for identifying new ARGs as they are potentially emerging in pathogens.

## 4.2 Paper II

When looking at the findings presented in previous studies, it becomes clear that our current understanding of the resistome is highly limited [74, 75, 140]. Indeed, we still lack fundamental knowledge about the genetic diversity of ARGs present in different environments that could potentially be acquired by pathogens in the future. This makes it difficult to anticipate new ARGs as they emerge in pathogens and, consequently, to develop effective strategies to manage the spread of new forms of resistance. In Paper II, *Latent antibiotic resistance genes are abundant, diverse, and mobile in human, animal, and environmental microbiomes*, our objective, therefore, was to provide a more complete overview of the abundance and diversity of ARGs in external and host-associated environments. To do this, we first constructed a large reference database encompassing both well-known (here denoted as "established") and putative (here denoted as "latent") ARGs from 17 different gene classes. In total, we included 572 established genes, which were obtained from the ResFinder database, as well as 23,502 latent genes, which were identified by analyzing 427,495 bacterial genomes with the fARGene software [74]. The abundance of these ARGs was then estimated in 10,744 metagenomic samples, representing 20 environment types (Figure 4.2).

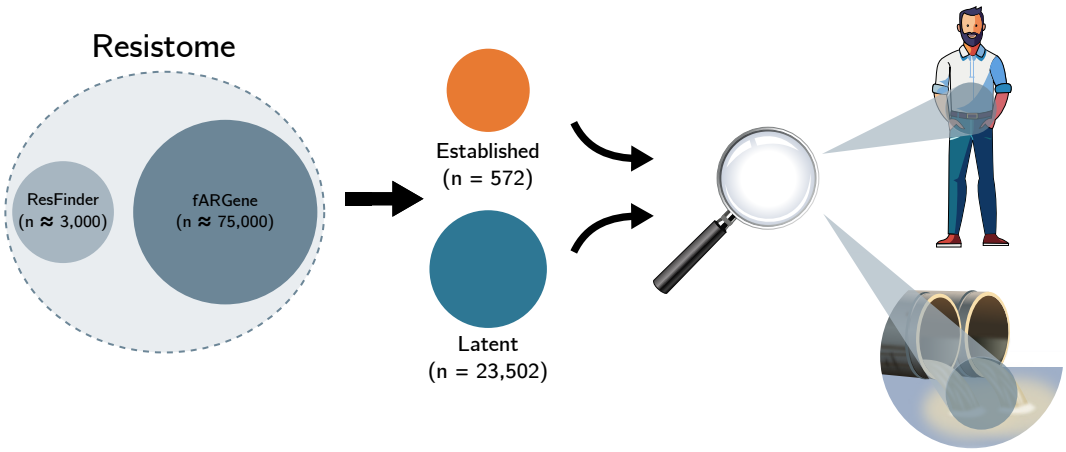



**Figure 4.2:** Overview of the analysis pipeline. To gain insight into the resistome, we separated ARGs into two distinct groups: the established ARGs, consisting of mobile ARGs that are already clinically relevant, and the latent ARGs, consisting of computationally predicted "new" genes. Each group was carefully curated to encompass genetically dissimilar ARGs. Subsequently, we searched for established and latent ARGs within an extensive metagenomic database spanning diverse environments.

When analyzing the estimated gene abundances across the environments,

we found stark contrasts between the abundance profiles associated with latent and established ARGs (Figure 4.3). Latent ARGs were generally more abundant than their established counterparts in all external environments except wastewater, while host-associated environments were found to contain a mix of both latent and established variants.

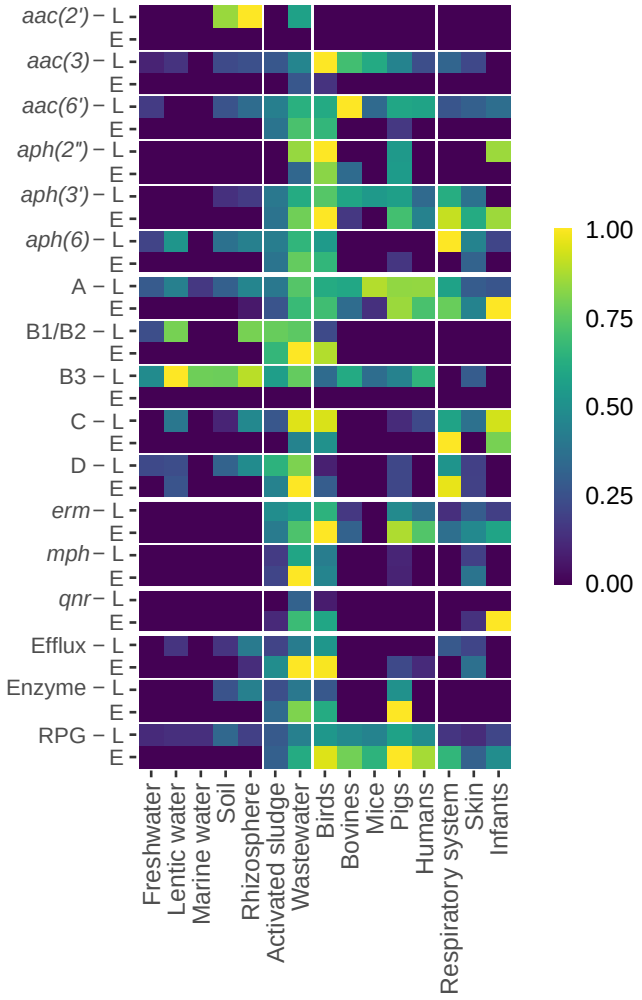



**Figure 4.3:** Relative abundance of latent and established ARGs divided by gene class and environment. Each gene class is represented by two rows: L for latent and E for established. The labels Birds, Bovines, Mice, Pigs, Humans, and Infants denote metagenomes from the corresponding digestive system. Respiratory system and Skin only include human samples. The color intensity reflects the gene- and environment-specific relative abundance, which was calculated based on the median of the relative log-transform abundance over all samples from the environment. To make the genes comparable, all values were normalized based on the environment with the highest abundance. RPG is short for ribosomal protection gene. From Inda-Díaz et al. 2023 (*Microbiome*. 2023 Mar 8;11(1):44). Licensed under CC-BY-4.0.

Based on the observed presence or absence of ARGs in different metagenomic samples, we then estimated the pan-resistome (i.e. the complete collection of ARGs present in any sample) and the core-resistome (i.e., the subset of ARGs that were consistently present across samples) of each environment. Here, we found that all pan-resistomes were dominated by latent ARGs, with the pan-resistomes of external environments on average showing greater genetic diversity than their host-associated counterparts. Furthermore, we found significant overlaps between the core-resistomes of human and animal digestive systems and wastewater, suggesting that the microbes that colonize these environments are subjected to similar selection pressures.

Finally, to investigate the potential mobility of the latent core-resistome, we extracted and annotated the genetic contexts of latent core-resistome ARGs from whole-genome sequencing data in a similar manner as described for Paper I. Here, we found that of the 29 latent ARGs that were included in the core-resistomes of at least two different environments, 48% were located close to a gene associated with MGEs, while 21% were co-localized with an established ARG. Taken together, the results of this study show that latent ARGs are ubiquitous in all analyzed environments, with a genetic diversity that far surpasses established variants, and have the ability to transfer horizontally within and between environments. Thus, latent ARGs also need to be considered in future studies to provide a more comprehensive view of the resistome and its implications for human health.

## 4.3 Paper III

A key finding from Paper II was the significant similarities between the core-resistomes of human gut and wastewater environments. Based on their frequent exposure to antimicrobial compounds and their taxonomic composition, these microbiomes have previously been suggested as hotspots for the spread of antibiotic resistance [30, 33]. In Paper III, *Community-promoted antibiotic resistance genes show increased dissemination among pathogens*, we therefore set out to elucidate the connection between the prevalence of ARGs in human gut and wastewater microbial communities, and their potential implications for human health.

We used the same basic approach as outlined for Paper II to estimate the prevalence of ARGs in metagenomes, however, we first updated our reference ARG database to include 720 established ARGs and 33,224 latent ARGs. The presence of each of these genes was estimated in a total of 6,664 metagenomic shotgun samples, 5,630 of which represented the human gut and 1,034 of

which represented wastewater. For each included ARG, we then calculated the proportion of samples from each environment in which the gene was present ($\geq 3$ matching reads). Based on these proportions, we then divided the ARGs into four categories; co-promoted ARGs, which were present in $\geq 5\%$ of samples from both environments, non-promoted ARGs, which were present in $< 5\%$ of samples from both environments (but not completely absent), and human gut (HG)-promoted and wastewater (WW)-promoted ARGs, both of which were present in $\geq 5\%$ of samples from one type of environment but not the other (Figure 4.4).

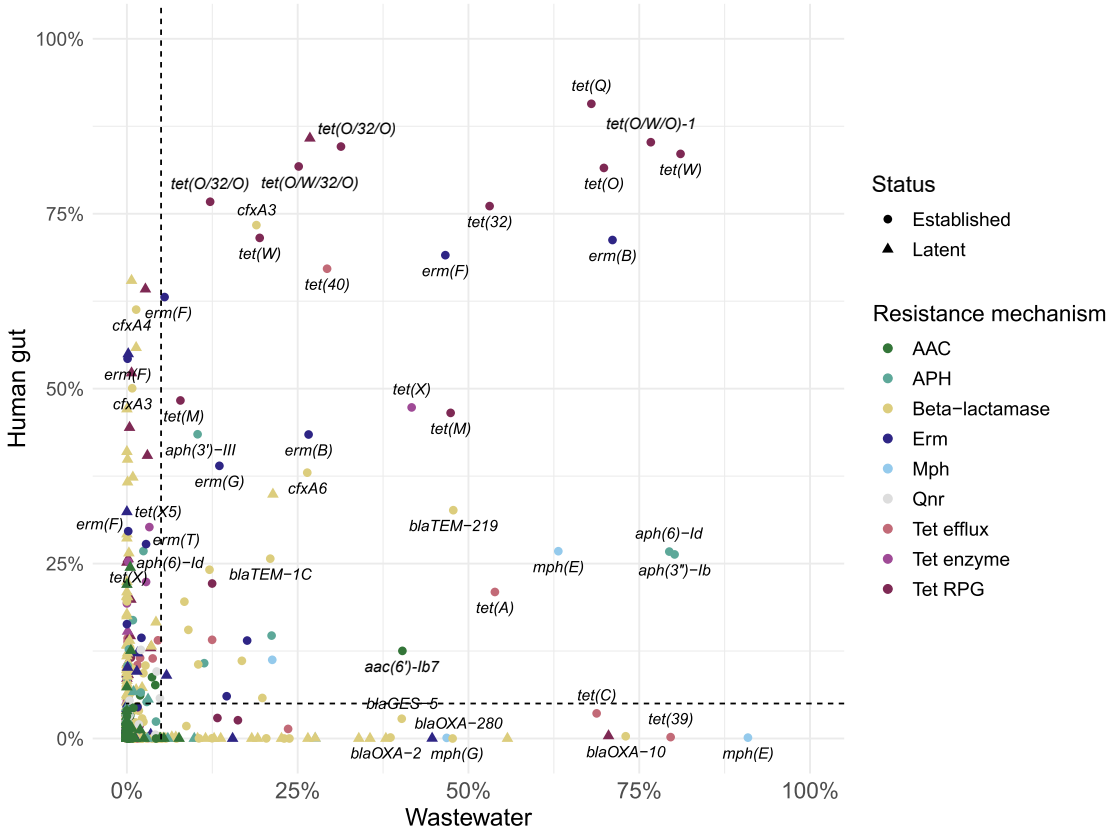



**Figure 4.4:** Prevalence of antibiotic resistance genes (ARGs) in human gut and wastewater metagenomic samples. Labels are included for established ARGs that were present in $\geq 25\%$ of samples from either environment. The dashed lines show the classification of the ARGs into four groups: Co-promoted (top right), human gut-promoted (top left), wastewater-promoted (bottom right), and non-promoted (bottom left). From Lund et al. 2025 (Preprint).

The co-promoted group mainly encompassed established ARGs with clinical significance, while the other groups were instead dominated by latent genes, though each category also included at least some established ARGs. By analyzing the bacterial hosts carrying ARGs belonging to different promotion categories, we found that the co-promoted genes were especially widespread. Indeed, these genes were significantly overrepresented among ARGs identified in multiple bacterial phyla, suggesting an increased potential for wide horizontal dissemination. By contrast, WW-promoted ARGs were overrepresented among ARGs identified in multiple classes within the same phylum (Pseudomonadota), showing that while these genes are promiscuous they are more taxonomically restricted. Together, co-promoted and WW-promoted ARGs were also overrepresented among established ARGs identified in multiple important bacterial pathogens.

When analyzing the genetic contexts of the ARGs from the four different categories, we found that established co-promoted ARGs were more frequently identified near broad host-range conjugative elements compared to HG- and WW-promoted ARGs. This might explain the higher propensity of co-promoted ARGs to transfer over long evolutionary distances. Moreover, we also estimated the genetic compatibility (nucleotide composition dissimilarity) between the established ARGs and genomes representing the included bacterial pathogens, as well as typical residents of the human gut and wastewater bacterial communities. Here, we again found that the co-promoted ARGs, on average, were genetically more similar to these genomes than other established ARGs. This suggests that these genes might be more easily assimilated by the members of these microbial communities, including pathogens, which might further facilitate their horizontal dissemination.

In summary, this paper represents a systematic investigation of the resistomes found in the human gut and wastewater microbiomes. By identifying the properties associated with genes promoted in different environments, we were able to shed light on the correlations between prevalence in microbial communities and transfer into common human pathogens. Together with papers I and II, this represents an unprecedented insight into the resistome in general, and the mobile resistome in particular.

## 4.4 Paper IV

The acquisition of foreign genetic material through HGT is one of the main ways by which antibiotic resistance is spreading. In a single transfer event, a cell can develop resistance to multiple antibiotics, which enables the rapid evolution

of multidrug-resistant pathogens under appropriate selection pressure [141]. Furthermore, it has been shown that pathogens have repeatedly recruited ARGs originating from evolutionarily distant bacteria [142]. Thus, the horizontal transfer of ARGs across long evolutionary distances has played a key part in the development of multi-resistance over time. However, while the negative impact of horizontal ARG transfer between evolutionarily divergent bacteria on human health is undeniable, the details regarding this process remain largely unknown. To overcome the threat posed by increasing antibiotic resistance, it is vital that we increase our knowledge about how ARGs move between different bacteria.

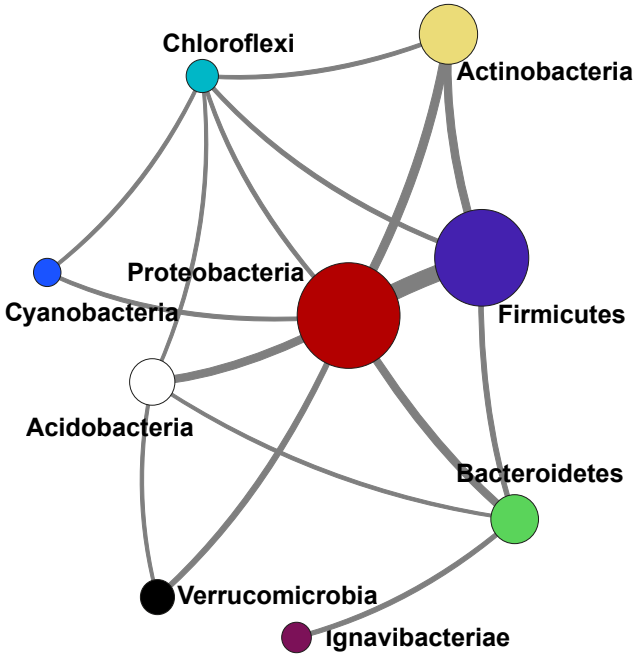



**Figure 4.5:** Network representation of the inter-phyla transfers between Proteobacteria, Firmicutes, Actinobacteria, Chloroflexi, Cyanobacteria, Acidobacteria, Verrucomicrobia, and Bacteroidetes. From Parras-Moltó et al. 2024 (Preprint).

In Paper IV, *The transfer of antibiotic resistance genes between evolutionarily distant bacteria*, we aimed to systematically analyze the transfer of ARGs between bacterial phyla by identifying the associated taxonomic patterns (i.e., which phyla are most frequently sharing ARGs) as well as the environments where transfers are most frequently occurring. To identify instances of inter-phyla ARG transfer, we implemented an algorithm based on phylogenetic analysis that identified discrepancies between the phylogenetic trees reconstructed from the predicted protein sequences from each gene class and the recorded host taxonomy (the details behind this approach are described in Chapter 3.3). This

algorithm was then applied to ARGs from 22 different gene classes predicted by fARGene [70] in almost half a million publicly available bacterial genomes, resulting in 661 identified inter-phyla transfers (IPTs). Here, Proteobacteria was the phylum most frequently engaging in IPT, followed by Firmicutes and Acidobacteria (Figure 4.5). Furthermore, our results revealed that Proteobacteria have played a central role in the IPT of all ARG classes except Erm 23S rRNA methyltransferases and tetracycline ribosomal protection genes (RPGs), which were instead more strongly associated with Firmicutes. When increasing the taxonomic resolution, we found Bacilli (Firmicutes) to be the class involved in the highest number of IPTs, mainly together with Gammaproteobacteria, where transfers mostly involved RPGs, and Epsilonproteobacteria, where transfers mostly involved aminoglycoside-modifying enzymes (Figure 4.6).

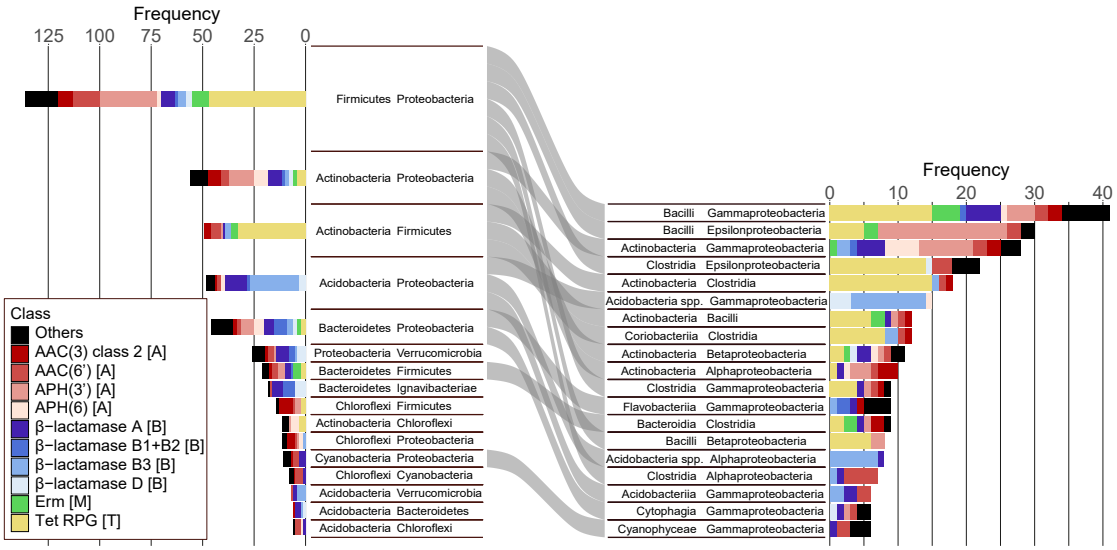



**Figure 4.6:** The most common transfers involving taxonomic classes from different bacterial phyla, stratified based on the class of the transferred ARG. From Parras-Moltó et al. 2024 (Preprint).

To investigate within what environments the observed IPTs were likely to have occurred, we extracted the reported isolation sources of the bacterial genomes carrying ARGs involved in IPTs. Here, we found that over half of the genomes for which this information was available were isolated from the human microbiome. Statistical analysis revealed that IPTs involving Mph macrolide 2'-phosphotransferases, class B1/B2 and D beta-lactamases, and tetracycline RPGs in particular were significantly overrepresented in the human microbiome. By contrast, IPTs involving aminoglycoside-modifying enzymes and class B3 beta-lactamases were significantly associated with external environments, including soil and water. Interestingly, analysis of sequence

similarities suggested that IPTs identified in the human microbiome were more recent than those identified in external environments. Taken together, our results provide new insights into the evolutionary process behind the accumulation of ARGs in bacteria, which has been key for the development of multidrug-resistant pathogens.

## 4.5 Paper V

In Paper V, *Genetic compatibility and ecological connectivity drive the dissemination of antibiotic resistance genes*, we further developed the methodology from Paper IV, with the aim of quantifying the extent to which different factors influence horizontal ARG transfer over large phylogenetic distances (Figure 4.7). First, we expanded the dataset to include $\sim 800,000$ bacterial genomes whose taxonomic annotations were of sufficient quality. These genomes were again screened for ARGs using fARGene, and instances where ARGs had undergone horizontal transfer were identified based on phylogenetic analysis. However, compared to the previous study, where we only studied inter-phylum transfers, we opted to instead identify transfers between host bacteria with at least an order-level taxonomic difference. In total, this analysis yielded 6,276 identified transfers between distantly related host pairs.

For each identified transfer, data was collected representing the genetic incompatibility of the ARG and its host genomes, as well as the estimated environmental co-occurrence of donor and recipient genomes in different environments. Briefly, genetic incompatibility was calculated based on the nucleotide composition dissimilarity (estimated as difference in 5mer distributions) between the taxonomically distant host genomes, as well as between the transferred ARG and host genomes, while co-occurrence was estimated by mapping the genomes involved in the identified transfers onto a large 16S metagenomic dataset including 20,816 metagenomes from five different environment types. This data was supplemented with information on the cell envelope composition of the involved bacterial taxa, as well as the proportional difference in size between host genomes, and used as input features to train random forest models for prediction of horizontal ARG transfer.

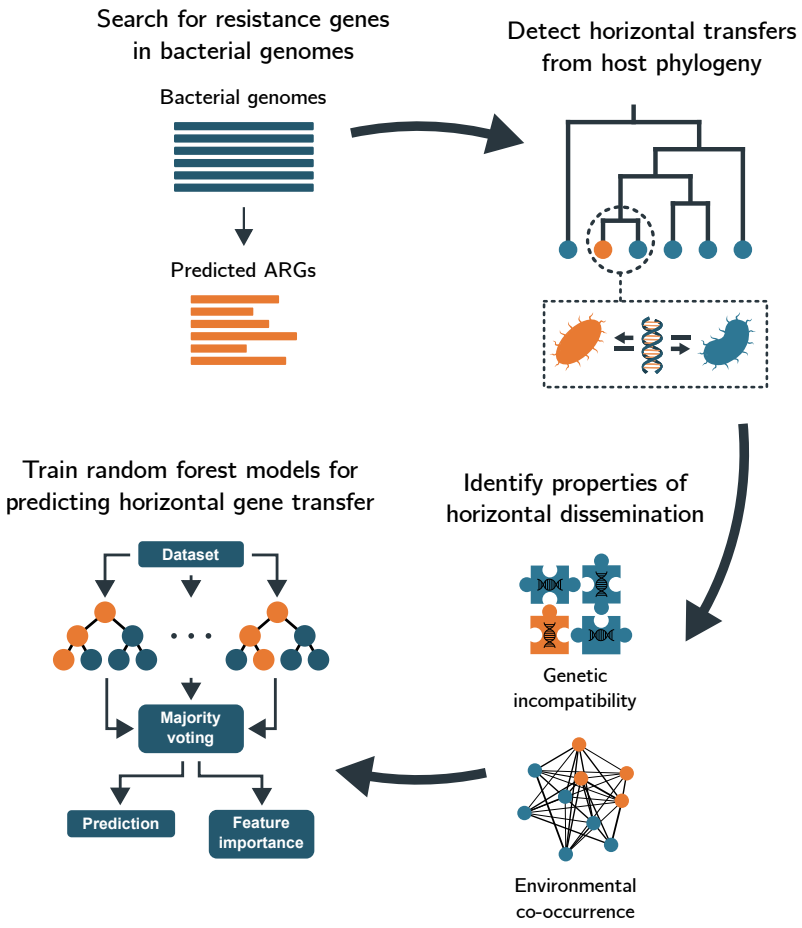



**Figure 4.7:** Overview of the analysis pipeline. Bacterial genomes were screened for ARGs, and phylogenetic trees were built from the identified sequences. Horizontal transfer was then inferred from the trees by detecting similar genes carried by evolutionarily distant hosts. For each identified transfer, data describing genetic incompatibility and co-occurrence in bacterial communities were collected and used to train random forest models, from which the most influential factors were identified. Adapted from Lund et al. 2025 (*Nature Communications*. 2025 Mar 16;16(1):2595). Licensed under CC-BY-4.0.

In total, eight models were created, including one general and seven models specific to different resistance mechanisms. Here, all models were trained using a positive dataset consisting of observed transfers and a negative dataset created by randomly permuting the leaves in the phylogenetic gene trees. When evaluated based on performance, the final models displayed mean areas under the receiver operating characteristic curve (AUROC) between 0.810–0.930, mean sensitivities between 0.803–0.902, and mean specificities between 0.710–0.874, showing that they were able to accurately identify most observed transfers while maintaining a low false positive rate.

By combining feature importance analysis with partial dependence analysis, we were able to identify the most influential factors for successful HGT of different ARGs, as well as assess whether different factors generally had a positive or negative effect on this process. Our results revealed that genetic incompatibility, i.e. the nucleotide composition dissimilarity between genomes and/or between genomes and ARGs, had the largest effect on all models (Figure 4.8), and negatively affected the likelihood of horizontal ARG transfer. Similarly, a pronounced negative effect was seen for hosts with different Gram staining profiles, with different resistance mechanisms favoring transfer between either Gram-negatives or Gram-positives. Finally, co-occurrence in any environment was found to generally have a positive impact on horizontal ARG transfer, however, co-occurrence in the human and wastewater microbiomes had the strongest effect on the likelihood of transfer.

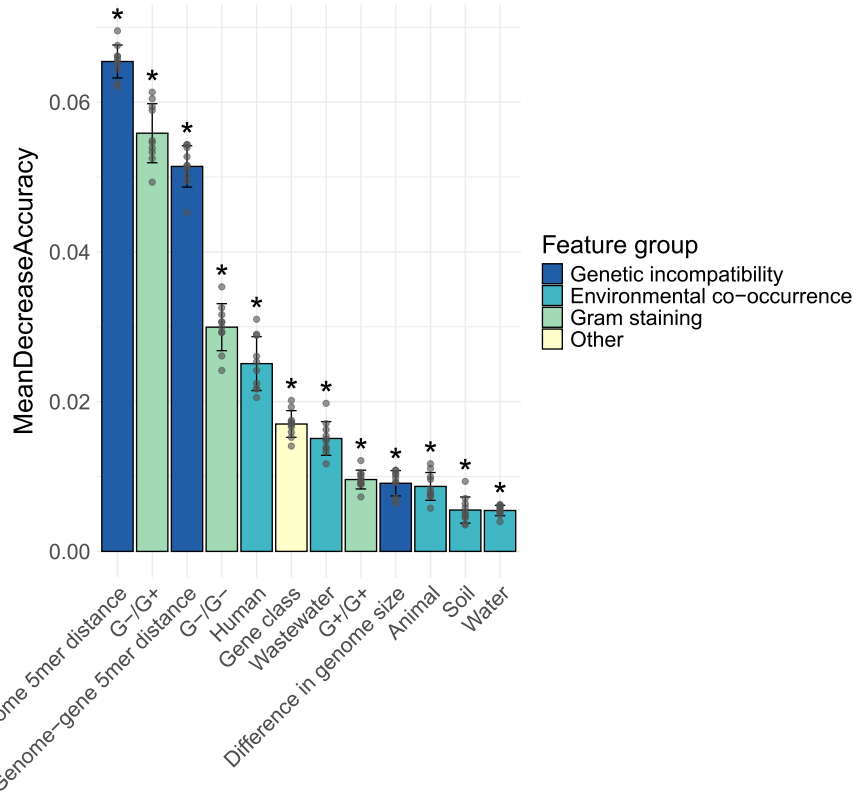



**Figure 4.8:** Relative importance of genetic and environmental factors for predicting the horizontal transfer of antibiotic resistance genes. The bars show the mean +/- SD of the importance of each factor to the accuracy of the general random forest model, based on all observed transfers ($n = 1,565$), over ten iterations. Permutation tests were used to generate a p-value for each factor and iteration. $^{*}P < 0.01$ across all model iterations. From Lund et al. 2025 (*Nature Communications*. 2025 Mar 16;16(1):2595). Licensed under CC-BY-4.0.

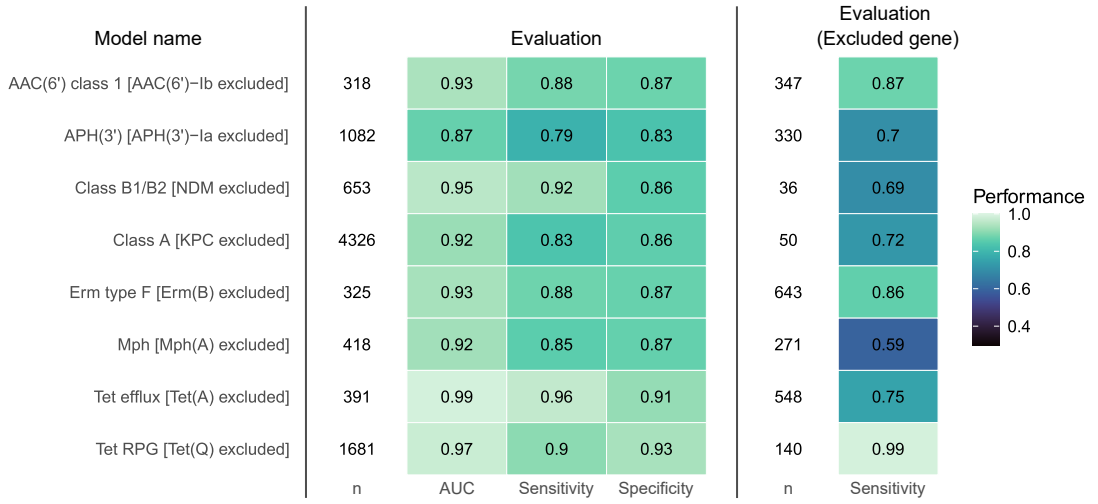
From network analysis of the observed co-occurrence patterns, we found that these were highly environment-specific. In particular, the human microbiome displayed the greatest diversity of co-occurring promiscuous bacterial hosts. By contrast, the observed co-occurrence of bacteria engaging in HGT in wastewater was more taxonomically restricted, being dominated by Pseudomonadota, but the co-occurrence levels were generally higher. Taken together, our findings show that the difference in genetic composition of host bacteria and their co-occurrence in microbial communities are two key factors that shape the dissemination of ARGs among environmental, commensal, and disease-causing bacteria. The results from this study also highlight the potential of predictive models for early detection of emerging resistance determinants.

## 4.6 Paper VI

The results from Paper V clearly show that machine-learning models can be trained to identify horizontal ARG transfer with high accuracy. In Paper VI, *Can we predict the spread of novel antibiotic resistance genes?*, we aimed to further refine the predictive power of these models and investigate the potential they have for predicting horizontal transfer of emerging ARGs. To do this, we first updated our dataset to encompass $\sim 1.6$ million bacterial genomes, which were screened for family-level horizontal ARG transfers. Using this expanded data, we were able to create gene class-specific models using the XGBoost framework, where we included additional features representing biological functions strongly correlated with either the presence or absence of each included gene class in bacterial genomes. The resulting 22 models generally displayed very high performance, showing that our updated approach improved the capacity to predict horizontal ARG transfer.

To simulate emerging resistance genes, we selected eight highly promiscuous ARGs, including the aminoglycoside modifying enzymes AAC(6')-Ib and APH(3')-Ia, the beta-lactamases NDM and KPC, the macrolide resistance genes *erm(B)* and *mph(A)*, and the tetracycline resistance genes *tet(A)* and *tet(B)*, and created new models where the transfers involving these ARGs had been excluded from the training data. These new models were evaluated based on their performance on test data (30% of the input data split randomly), but also in their ability to correctly classify the observations excluded from their respective input data (Figure 4.9). While the performance on test data was high for every model, the ability to predict the excluded genes varied greatly. Interestingly, we noted that there was generally a higher degree of difficulty associated with predicting inter-phyla transfers of an unseen ARG compared to transfers within the same bacterial phylum. Illustrating this, when inter-

phyla transfers were removed from consideration, the Mph [Mph(A) excluded] model was able to correctly classify 95% of Mph(A) transfers (as opposed to 59% when phylum-level transfers were included).




| Model name | n | AUC | Sensitivity | Specificity | n | Sensitivity |
|---|---|---|---|---|---|---|
| AAC(6') class 1 [AAC(6')−Ib excluded] | 318 | 0.93 | 0.88 | 0.87 | 347 | 0.87 |
| APH(3') [APH(3')−Ia excluded] | 1082 | 0.87 | 0.79 | 0.83 | 330 | 0.7 |
| Class B1/B2 [NDM excluded] | 653 | 0.95 | 0.92 | 0.86 | 36 | 0.69 |
| Class A [KPC excluded] | 4326 | 0.92 | 0.83 | 0.86 | 50 | 0.72 |
| Erm type F [Erm(B) excluded] | 325 | 0.93 | 0.88 | 0.87 | 643 | 0.86 |
| Mph [Mph(A) excluded] | 418 | 0.92 | 0.85 | 0.87 | 271 | 0.59 |
| Tet efflux [Tet(A) excluded] | 391 | 0.99 | 0.96 | 0.91 | 548 | 0.75 |
| Tet RPG [Tet(Q) excluded] | 1681 | 0.97 | 0.9 | 0.93 | 140 | 0.99 |

**Figure 4.9:** Performance of the selected models. For each model, transfers involving a selected reference gene were excluded from the dataset before training. After training the model, it was first evaluated based on test data (without the excluded reference genes), and then again based on the excluded set of transfers involving the reference gene. The left heatmap shows the results from the first round of evaluation, including the area under the receiver operating characteristic curve (AUC), sensitivity, and specificity, as well as the number of transfers in the test data (n) for each model. The right heatmap shows the sensitivity produced from the second round of evaluation, as well as the number of excluded transfers (n) for each model. From Lund et al. 2025 (Manuscript).

Next, we wanted to assess the predictive performance of each model when applied to an independent dataset. To create such a dataset, we first selected representative genomes for each of the 500 most common bacterial species in the NCBI Assembly database. We then generated all possible pairs of species whose taxonomic distance was at least at the family level and computed the same features for these pairs as we previously did for the observed horizontal transfers. Finally, we used each of the models described in the previous paragraph to assess the horizontal transfer potential of their respective excluded ARG for each combination of bacterial families. The results showed that all models predicted a wide range of potential transfer events. Here, a substantial amount of predicted connections involved documented hosts of the relevant ARGs, which were validated via the data and a literature review. Each also produced a wide range of novel predictions, a large portion of which involved less well-studied bacterial taxa. Thus, while these predicted novel transfers

likely included some false positives, the extent of this was difficult to assess. Taken together, our results indicate that it is indeed possible to use machine learning to predict the dissemination of emerging ARGs, at least to a certain degree, but also that further development is needed to improve the reliability of the models.

# 5 Conclusions

Antibiotic resistance represents one of the greatest threats to the future of human health globally. To overcome this, we need to increase our knowledge about how resistance is spreading, which is in large part driven by the horizontal dissemination of antibiotic resistance genes (ARGs). Fortunately, recent years have seen a vast expansion of biological sequence repositories, and this data provides opportunities for obtaining new insights about the antibiotic resistance phenomenon.

In this thesis, data-driven methods have been used to extend our knowledge about the diversity of ARGs carried by bacteria, the prevalence of these genes in different environments, and how they are disseminated among microbial communities. The insights provided by Papers I–III, which represent an unprecedented characterization of the resistome (or complete collection of ARGs in bacteria), allow for a greater understanding of the environmental origins and evolution of different resistance genes. Moreover, Papers I and II highlight many examples of previously unknown (latent) ARGs that show evidence of emerging in human pathogens. When considering that some of these latent ARGs were shown to provide clinical levels of resistance in an *E. coli* host (Paper I) and that some were identified in a range of different environments (Paper II), it is clear that the latent resistome encompasses many genes that likely constitute future clinical threats. As sequencing-based methods become more widely used for molecular diagnostics of clinical infections, it is crucial that antibiotic resistance profiling goes beyond the identification of only established ARGs. To ensure accurate results and, by extension, appropriate treatment, latent ARGs also need to be considered.

The presented findings also highlight the need for improved monitoring of ARGs across environments. Indeed, Paper III shows that prevalence in the human gut and wastewater microbiomes is strongly connected to widespread dissemination among bacterial pathogens. Thus, it is essential to be aware of new ARGs that become promoted in these environments, as such genes will

likely have negative implications for human health. Exemplifying this, many of the potentially emerging genes identified in Paper II were associated with either human or wastewater microbial communities, though not both, suggesting that these ARGs may not yet have achieved widespread dissemination among pathogens. Nevertheless, many of these genes appear to have been mobilized and transferred into pathogens undetected. This shows limitations in contemporary surveillance programs for monitoring antibiotic resistance, and highlights the potential of computational screening to detect potentially dangerous new ARGs before they spread widely – something traditional microbiological methods have largely not been successful with. Early identification of these genes will allow us to anticipate them as they appear in clinical settings and react accordingly.

To better evaluate the risks associated with novel ARGs, it is crucial to estimate their potential for dissemination through horizontal gene transfer (HGT). The findings presented in Papers IV–VI, which represent a detailed investigation into the patterns underlying the horizontal transfer of ARGs, increase the knowledge about how ARGs are disseminated, which could facilitate such risk assessment. Mainly, Paper IV shows that Proteobacteria act as a central hub for inter-phyla transfer of ARGs from most gene classes. Moreover, strong associations were identified between recent inter-phyla transfers and bacteria isolated from the human microbiome. This aligns with the findings from Papers II and III, again suggesting that these bacterial communities play an important role in the dissemination of mobile ARGs. The importance of the human and wastewater microbiomes is further supported by Paper V, where co-occurrence in these environments was shown to increase the likelihood of horizontal ARG transfer between bacterial orders. Thus, Papers II–V all highlight anthropogenic microbiomes as high-risk environments for horizontal ARGs dissemination, where an increased co-occurrence of potential host bacteria strongly increases the likelihood of successful horizontal ARG transfer. Consequently, while external environments should not be overlooked, the human gut and wastewater resistomes, including latent ARGs, should be of the highest concern when designing strategies to combat the antibiotic resistance crisis. The findings from Paper V, however, reveal that the factor contributing most strongly towards successful horizontal ARG transfer was genetic incompatibility — or nucleotide composition dissimilarity — between the genes and genomes involved in each transfer. The influence of genetic (in)compatibility is also described in Paper III, highlighting this as a key component shaping the dissemination of ARGs among environmental, commensal, and/or pathogenic bacteria, with mobile genes likely having an increased host-range the less they deviate from the genetic makeup of their potential hosts.

In Paper V, the model structure was further refined and used to identify novel

ARGs with high dissemination potential. The results show that, while this is a highly difficult task, it is not infeasible. Indeed, all presented models were able to predict a majority of transfers involving simulated emerging ARGs, showing that machine learning-based prediction of horizontal ARG dissemination is a viable strategy to anticipate upcoming threats. Moreover, many of the predictions produced from an independently generated dataset were not present in the data but were supported by the literature, showing that the models are able to generalize beyond the training data. However, many of these predictions could not be explained based on the available information, suggesting at least some degree of false positives, and showing that further development is needed before such models can support decision-making in clinical settings. Nevertheless, the potential of data-driven methods for anticipating emerging ARGs is clear, and, with further refinement, machine learning models could become an important asset in the fight against antibiotic resistance.

## 5.1 Future research

This thesis has shown in several ways how data-driven methods constitute an important asset for combating the spread of antibiotic resistance. Indeed, by identifying potentially emerging ARGs in whole-genome sequencing data, analyzing the resistome to identify high-risk environments for ARG dissemination using metagenomics, elucidating the evolutionary history of horizontally transferred ARGs using phylogenetics, and identifying factors that strongly influence the dissemination of ARGs using machine learning, this work demonstrates the diverse applications of data-driven methods in this context. As the amount of data generated is only expected to grow in the coming years, new methods must be developed that take full advantage of the available information to extract novel biological insights. Here, the recent popularization of artificial intelligence frameworks like large language models presents an exciting opportunity, as these could potentially help us better track mobile gene sequences across genomes. Moreover, different algorithms or approaches could prove to be a better fit for the application pioneered in Paper VI, which needs to be explored in future research. In summary, the era of data that we find ourselves in today presents an unprecedented opportunity to leverage this information for overcoming problems facing human health. Therefore, it is important that data-driven methods are developed that can supplement traditional microbiological methods for monitoring the spread of antibiotic resistance and for developing strategies to limit the spread of emerging ARGs.

# Bibliography


[1] Matthew I Hutchings, Andrew W Truman, and Barrie Wilkinson. Antibiotics: past, present and future. *Current opinion in microbiology*, 51:72–80, 2019.

[2] Kathrin I Mohr. History of antibiotics research. *How to Overcome the Antibiotic Crisis*, pages 237–272, 2016.

[3] Anthony RM Coates, Gerry Halls, and Yanmin Hu. Novel classes of antibiotics or more of the same? *British journal of pharmacology*, 163(1): 184–194, 2011.

[4] Amin Talebi Bezmin Abadi, Albert A Rizvanov, Thomas Haertlé, and Nataliya L Blatt. World health organization report: current crisis of antibiotic resistance. *BioNanoScience*, 9(4):778–788, 2019.

[5] Alasdair MacGowan and Emily Macnaughton. Antibiotic resistance. *Medicine*, 45(10):622–628, 2017.

[6] Georgina Cox and Gerard D Wright. Intrinsic antibiotic resistance: mechanisms, origins, challenges and solutions. *International Journal of Medical Microbiology*, 303(6-7):287–292, 2013.

[7] Lucía Fernández, W James Gooderham, Manjeet Bains, Joseph B McPhee, Irith Wiegand, and Robert EW Hancock. Adaptive resistance to the "last hope" antibiotics polymyxin b and colistin in pseudomonas aeruginosa is mediated by the novel two-component regulatory system parr-pars. *Antimicrobial agents and chemotherapy*, 54(8):3372–3382, 2010.

[8] Santiago Sandoval-Motta and Maximino Aldana. Adaptive resistance to antibiotics in bacteria: a systems biology perspective. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 8(3):253–267, 2016.

[9] Nicole A Lerminiaux and Andrew DS Cameron. Horizontal transfer of antibiotic resistance genes in clinical environments. *Canadian journal of microbiology*, 65(1):34–44, 2019.

[10] Liam S Redgrave, Sam B Sutton, Mark A Webber, and Laura JV Piddock. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. *Trends in microbiology*, 22(8):438–445, 2014.

[11] Daniel J Rankin, Eduardo PC Rocha, and Sam P Brown. What traits are carried on mobile genetic elements, and why? *Heredity*, 106(1):1–10, 2011.

[12] Jose M Munita and Cesar A Arias. Mechanisms of antibiotic resistance. *Microbiology spectrum*, 4(2):4–2, 2016.

[13] Valeria Bortolaia, Rolf S Kaas, Etienne Ruppe, Marilyn C Roberts, Stefan Schwarz, Vincent Cattoir, Alain Philippon, Rosa L Allesoe, Ana Rita Rebelo, Alfred Ferrer Florensa, et al. Resfinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy*, 75(12): 3491–3500, 2020.

[14] Brian P Alcock, William Huynh, Romeo Chalil, Keaton W Smith, Amogelang R Raphenya, Mateusz A Wlodarski, Arman Edalatmand, Aaron Petkau, Sohaib A Syed, Kara K Tsang, et al. Card 2023: expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic Acids Research*, 2022.

[15] Xian-Zhi Li, Patrick Plésiat, and Hiroshi Nikaido. The challenge of efflux-mediated antibiotic resistance in gram-negative bacteria. *Clinical microbiology reviews*, 28(2):337–418, 2015.

[16] George P Dinos. The macrolide antibiotic renaissance. *British journal of pharmacology*, 174(18):2967–2983, 2017.

[17] Timothy Palzkill. Metallo-$\beta$-lactamase structure and function. *Annals of the New York Academy of Sciences*, 1277(1):91–104, 2013.

[18] Jessica Blair, Mark A Webber, Alison J Baylay, David O Ogbolu, and Laura JV Piddock. Molecular mechanisms of antibiotic resistance. *Nature reviews microbiology*, 13(1):42–51, 2015.

[19] Stefan Ebmeyer, Erik Kristiansson, and D G Joakim Larsson. A framework for identifying the recent origins of mobile antibiotic resistance genes. *Communications biology*, 4(1):1–10, 2021.

[20] Dongchang Sun, Katy Jeannot, Yonghong Xiao, and Charles W Knapp. Horizontal gene transfer mediated bacterial antibiotic resistance. *Frontiers in microbiology*, 10:1933, 2019.

[21] Sally R Partridge, Stephen M Kwong, Neville Firth, and Slade O Jensen. Mobile genetic elements associated with antimicrobial resistance. *Clinical microbiology reviews*, 31(4):10–1128, 2018.

[22] Mislav Acman, Ruobing Wang, Lucy van Dorp, Liam P Shaw, Qi Wang, Nina Luhmann, Yuyao Yin, Shijun Sun, Hongbin Chen, Hui Wang, et al. Role of mobile genetic elements in the global dissemination of the carbapenem resistance gene bla ndm. *Nature communications*, 13(1):1131, 2022.

[23] Laura S Frost, Raphael Leplae, Anne O Summers, and Ariane Toussaint. Mobile genetic elements: the agents of open source evolution. *Nature Reviews Microbiology*, 3(9):722–732, 2005.

[24] Christopher M Thomas and Kaare M Nielsen. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, 3(9):711–721, 2005.

[25] Kristin Hegstad, Haima Mylvaganam, Jessin Janice, Ellen Josefsen, Audun Sivertsen, and Dagfinn Skaare. Role of horizontal gene transfer in the development of multidrug resistance in haemophilus influenzae. *Msphere*, 5(1):e00969–19, 2020.

[26] Brian J Arnold, I-Ting Huang, and William P Hanage. Horizontal gene transfer and adaptive evolution in bacteria. *Nature Reviews Microbiology*, 20(4):206–218, 2022.

[27] Kimihiro Abe, Nobuhiko Nomura, and Satoru Suzuki. Biofilms: hot spots of horizontal gene transfer (hgt) in aquatic environments, with a focus on a new hgt mechanism. *FEMS microbiology ecology*, 96(5):fiaa031, 2020.

[28] Gyanendra P Dubey and Sigal Ben-Yehuda. Intercellular nanotubes mediate bacterial communication. *Cell*, 144(4):590–600, 2011.

[29] Pavol Bárdy, Tibor Füzik, Dominik Hrebík, Roman Pantček, John Thomas Beatty, and Pavel Plevka. Structure and mechanism of dna delivery of a gene transfer agent. *Nature Communications*, 11(1):3034, 2020.

[30] D G Joakim Larsson and Carl-Fredrik Flach. Antibiotic resistance in the environment. *Nature Reviews Microbiology*, 20(5):257–269, 2022.

[31] Julie Perry, Nicholas Waglechner, and Gerard Wright. The prehistory of antibiotic resistance. *Cold Spring Harbor perspectives in medicine*, 6(6): a025197, 2016.

[32] Gerard D Wright. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nature Reviews Microbiology*, 5(3):175–186, 2007.

[33] Chandan Pal, Johan Bengtsson-Palme, Erik Kristiansson, and D G Joakim Larsson. The structure and diversity of human, animal and environmental resistomes. *Microbiome*, 4(1):1–15, 2016.

[34] Mathieu Groussin, Mathilde Poyet, Ainara Sistiaga, Sean M Kearney, Katya Moniz, Mary Noel, Jeff Hooker, Sean M Gibbons, Laure Segurel, Alain Froment, et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell*, 184(8):2053–2067, 2021.

[35] Michael R Gillings. Evolutionary consequences of antibiotic use for the resistome, mobilome and microbial pangenome. *Frontiers in microbiology*, 4:4, 2013.

[36] Kristoffer Forslund, Shinichi Sunagawa, Jens Roat Kultima, Daniel R Mende, Manimozhiyan Arumugam, Athanasios Typas, and Peer Bork. Country-specific antibiotic use practices impact the human gut resistome. *Genome research*, 23(7):1163–1169, 2013.

[37] Molly K Gibson, Kevin J Forsberg, and Gautam Dantas. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*, 9(1):207–216, 2015.

[38] Yasmin Neves Vieira Sabino, Mateus Ferreira Santana, Linda Boniface Oyama, Fernanda Godoy Santos, Ana Júlia Silva Moreira, Sharon Ann Huws, and Hilário Cuquetto Mantovani. Characterization of antibiotic resistance genes in the species of the rumen microbiota. *Nature communications*, 10(1):5252, 2019.

[39] Ross S McInnes, Gregory E McCallum, Lisa E Lamberte, and Willem van Schaik. Horizontal transfer of antibiotic resistance genes in the human gut microbiome. *Current opinion in microbiology*, 53:35–43, 2020.

[40] Antti Karkman, Thi Thuy Do, Fiona Walsh, and Marko PJ Virta. Antibiotic-resistance genes in waste water. *Trends in microbiology*, 26 (3):220–228, 2018.

[41] Jinlyung Choi, Fan Yang, Ramunas Stepanauskas, Erick Cardenas, Aaron Garoutte, Ryan Williams, Jared Flater, James M Tiedje, Kirsten S Hofmockel, Brian Gelder, et al. Strategies to improve reference databases for soil microbiomes. *The ISME journal*, 11(4):829–834, 2017.

[42] Francis Crick. Central dogma of molecular biology. *Nature*, 227(5258): 561–563, 1970.

[43] Jay Shendure, Shankar Balasubramanian, George M Church, Walter Gilbert, Jane Rogers, Jeffery A Schloss, and Robert H Waterston. Dna sequencing at 40: past, present and future. *Nature*, 550(7676):345–353, 2017.

[44] Robert W Holley, Jean Apgar, George A Everett, James T Madison, Mark Marquisee, Susan H Merrill, John Robert Penswick, and Ada Zamir. Structure of a ribonucleic acid. *Science*, 147(3664):1462–1465, 1965.

[45] Robert D Fleischmann, Mark D Adams, Owen White, Rebecca A Clayton, Ewen F Kirkness, Anthony R Kerlavage, Carol J Bult, Jean-Francois Tomb, Brian A Dougherty, Joseph M Merrick, et al. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *science*, 269 (5223):496–512, 1995.

[46] J Craig Venter, Mark D Adams, Eugene W Myers, Peter W Li, Richard J Mural, Granger G Sutton, Hamilton O Smith, Mark Yandell, Cheryl A Evans, Robert A Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[47] Erwin L Van Dijk, Hélène Auger, Yan Jaszczyszyn, and Claude Thermes. Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426, 2014.

[48] Taishan Hu, Nilesh Chitnis, Dimitri Monos, and Anh Dinh. Next-generation sequencing technologies: An overview. *Human Immunology*, 82(11):801–811, 2021.

[49] Paul A Kitts, Deanna M Church, Françoise Thibaud-Nissen, Jinna Choi, Vichet Hem, Victor Sapojnikov, Robert G Smith, Tatiana Tatusova, Charlie Xiang, Andrey Zherikov, et al. Assembly: a resource for assembled genomes at ncbi. *Nucleic acids research*, 44(D1):D73–D80, 2016.

[50] Manish Boolchandani, Alaric W D'Souza, and Gautam Dantas. Sequencing-based methods and resources to study antimicrobial resistance. *Nature Reviews Genetics*, 20(6):356–370, 2019.

[51] Jean-Christophe Lagier, Sophie Edouard, Isabelle Pagnier, Oleg Mediannikov, Michel Drancourt, and Didier Raoult. Current and past strategies for bacterial culture in clinical microbiology. *Clinical microbiology reviews*, 28(1):208–236, 2015.

[52] Attila Bodor, Naila Bounedjoum, György Erik Vincze, Ágnes Erdeiné Kis, Krisztián Laczi, Gábor Bende, Árpád Szilágyi, Tamás Kovács, Katalin Perei, and Gábor Rákhely. Challenges of unculturable bacteria: environmental perspectives. *Reviews in Environmental Science and Bio/Technology*, 19:1–22, 2020.

[53] Can Su, Liping Lei, Yanqing Duan, Ke-Qin Zhang, and Jinkui Yang. Culture-independent methods for studying environmental microorganisms: methods, application, and perspective. *Applied microbiology and biotechnology*, 93:993–1003, 2012.

[54] Christopher Quince, Alan W Walker, Jared T Simpson, Nicholas J Loman, and Nicola Segata. Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833–844, 2017.

[55] Bo Yang, Yong Wang, and Pei-Yuan Qian. Sensitivity and correlation of hypervariable regions in 16s rrna genes in phylogenetic analysis. *BMC bioinformatics*, 17:1–8, 2016.

[56] Mark Blaxter, Jenna Mann, Tom Chapman, Fran Thomas, Claire Whitton, Robin Floyd, and Eyualem Abebe. Defining operational taxonomic units using dna barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1462):1935–1943, 2005.

[57] Benjamin J Callahan, Paul J McMurdie, and Susan P Holmes. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME journal*, 11(12):2639–2643, 2017.

[58] Patrick D Schloss. Amplicon sequence variants artificially split bacterial genomes into separate clusters. *Msphere*, 6(4):10–1128, 2021.

[59] Shashank Gupta, Martin S Mortensen, Susanne Schjørring, Urvish Trivedi, Gisle Vestergaard, Jakob Stokholm, Hans Bisgaard, Karen A Krogfelt, and Søren J Sørensen. Amplicon sequencing provides more accurate microbiome information in healthy children compared to culturing. *Communications biology*, 2(1):291, 2019.

[60] Structure, function and diversity of the healthy human microbiome. *nature*, 486(7402):207–214, 2012.

[61] Shinichi Sunagawa, Silvia G Acinas, Peer Bork, Chris Bowler, Damien Eveillard, Gabriel Gorsky, Lionel Guidi, Daniele Iudicone, Eric Karsenti, et al. Tara oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, 18(8):428–445, 2020.

[62] Johan Bengtsson-Palme, D G Joakim Larsson, and Erik Kristiansson. Using metagenomics to investigate human and environmental resistomes. *Journal of Antimicrobial Chemotherapy*, 72(10):2690–2703, 2017.

[63] Kristoffer Forslund, Shinichi Sunagawa, Luis P Coelho, and Peer Bork. Metagenomic insights into the human gut resistome and the forces that shape it. *Bioessays*, 36(3):316–329, 2014.

[64] Val F Lanza, Fernando Baquero, José Luís Martínez, Ricardo Ramos-Ruíz, Bruno González-Zorn, Antoine Andremont, Antonio Sánchez-Valenzuela, Stanislav Dusko Ehrlich, Sean Kennedy, Etienne Ruppé, et al. In-depth resistome analysis by targeted metagenomics. *Microbiome*, 6: 1–14, 2018.

[65] Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, et al. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic acids research*, 51(D1): D753–D759, 2023.

[66] Alfonso J Alanis. Resistance to antibiotics: are we in the post-antibiotic era? *Archives of medical research*, 36(6):697–705, 2005.

[67] Robert C Moellering Jr. Ndm-1—a cause for worldwide concern. *New England Journal of Medicine*, 363(25):2377–2379, 2010.

[68] Huiyan Ye, Yihui Li, Zhencui Li, Rongsui Gao, Han Zhang, Ronghui Wen, George F Gao, Qinghua Hu, and Youjun Feng. Diversified mcr-1-harbouring plasmid reservoirs confer resistance to colistin in human gut microbiota. *MBio*, 7(2):e00177–16, 2016.

[69] Ea Zankari, Henrik Hasman, Salvatore Cosentino, Martin Vestergaard, Simon Rasmussen, Ole Lund, Frank M Aarestrup, and Mette Voldby Larsen. Identification of acquired antimicrobial resistance genes. *Journal of antimicrobial chemotherapy*, 67(11):2640–2644, 2012.

[70] Fanny Berglund, Tobias Österlund, Fredrik Boulund, Nachiket P Marathe, D G Joakim Larsson, and Erik Kristiansson. Identification and reconstruction of novel antibiotic resistance genes from metagenomes. *Microbiome*, 7(1):1–14, 2019.

[71] Lubna Maryam, Salman Sadullah Usmani, and Gajendra PS Raghava. Computational resources in the management of antibiotic resistance: speeding up drug discovery. *Drug Discovery Today*, 26(9):2138–2151, 2021.

[72] Sean R. Eddy. Profile hidden markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763, 1998.

[73] Fredrik Boulund, Fanny Berglund, Carl-Fredrik Flach, Johan Bengtsson-Palme, Nachiket P Marathe, D G Joakim Larsson, and Erik Kristiansson. Computational discovery and functional validation of novel fluoroquinolone resistance genes in public metagenomic data sets. *BMC genomics*, 18(1):1–9, 2017.

[74] Fanny Berglund, Nachiket P Marathe, Tobias Österlund, Johan Bengtsson-Palme, Stathis Kotsakis, Carl-Fredrik Flach, D G Joakim Larsson, and Erik Kristiansson. Identification of 76 novel b1 metallo-$\beta$-lactamases through large-scale screening of genomic and metagenomic data. *Microbiome*, 5(1):1–13, 2017.

[75] Fanny Berglund, Maria-Elisabeth Böhm, Anton Martinsson, Stefan Ebmeyer, Tobias Österlund, Anna Johnning, D G Joakim Larsson, and Erik Kristiansson. Comprehensive screening of genomic and metagenomic data reveals a large diversity of tetracycline resistance genes. *Microbial genomics*, 6(11), 2020.

[76] Gustavo Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S Heath, Peter Vikesland, and Liqing Zhang. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):1–15, 2018.

[77] Etienne Ruppé, Amine Ghozlane, Julien Tap, Nicolas Pons, Anne-Sophie Alvarez, Nicolas Maziers, Trinidad Cuesta, Sara Hernando-Amado, Irene Clares, Jose Luís Martínez, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nature microbiology*, 4(1): 112–123, 2019.

[78] Victor Kunin, Alex Copeland, Alla Lapidus, Konstantinos Mavromatis, and Philip Hugenholtz. A bioinformatician's guide to metagenomics. *Microbiology and molecular biology reviews*, 72(4):557–578, 2008.

[79] Jay S Ghurye, Victoria Cepeda-Espinoza, and Mihai Pop. Metagenomic assembly: overview, challenges and applications. *The Yale journal of biology and medicine*, 89(3):353, 2016.

[80] Johan Bengtsson-Palme, Fredrik Boulund, Jerker Fick, Erik Kristiansson, and DG Joakim Larsson. Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in india. *Frontiers in microbiology*, 5:648, 2014.

[81] Jian Ye, Scott McGinnis, and Thomas L Madden. Blast: improvements for better sequence analysis. *Nucleic acids research*, 34(suppl_2):W6–W9, 2006.

[82] Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.

[83] Yao Pei, Marcus Ho-Hin Shum, Yunshi Liao, Vivian W Leung, Yu-Nong Gong, David K Smith, Xiaole Yin, Yi Guan, Ruibang Luo, Tong Zhang, et al. Argnet: using deep neural networks for robust identification and classification of antibiotic resistance genes from sequences. *Microbiome*, 12(1):84, 2024.

[84] Xiaole Yin, Xiawan Zheng, Liguan Li, An-Ni Zhang, Xiao-Tao Jiang, and Tong Zhang. Args-oap v3. 0: Antibiotic-resistance gene database curation and analysis pipeline optimization. *Engineering*, 27:234–241, 2023.

[85] Will PM Rowe and Martyn D Winn. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics*, 34(21):3601–3608, 2018.

[86] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[87] Mark Wilkinson, James O McInerney, Robert P Hirt, Peter G Foster, and T Martin Embley. Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends in ecology & evolution*, 22(3):114–115, 2007.

[88] John P Huelsenbeck, Jonathan P Bollback, and Amy M Levine. Inferring the root of a phylogenetic tree. *Systematic biology*, 51(1):32–43, 2002.

[89] Paschalia Kapli, Ziheng Yang, and Maximilian J Telford. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, 21(7):428–444, 2020.

[90] Naruya Saitou and Masatoshi Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425, 1987.

[91] Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature reviews genetics*, 13(5):303–314, 2012.

[92] Joseph Felsenstein. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376, 1981.

[93] E Michu. A short guide to phylogeny reconstruction. *Plant Soil and Environment*, 53(10):442, 2007.

[94] Joseph Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology*, 22(3):240–249, 1973.

[95] Mark E Siddall and Arnold G Kluge. Probabilism and phylogenetic inference. *Cladistics*, 13(4):313–336, 1997.

[96] John P Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P Bollback. Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550):2310–2314, 2001.

[97] Mark Holder and Paul O Lewis. Phylogeny estimation: traditional and bayesian approaches. *Nature reviews genetics*, 4(4):275–284, 2003.

[98] Yu K Mo, Matthew W Hahn, and Megan L Smith. Applications of machine learning in phylogenetics. *Molecular Phylogenetics and Evolution*, 196:108066, 2024.

[99] Hatch W Stokes and Michael R Gillings. Gene flow, mobile genetic elements and the recruitment of antibiotic resistance genes into gram-negative pathogens. *FEMS microbiology reviews*, 35(5):790–819, 2011.

[100] Matt Ravenhall, Nives Škunca, Florent Lassalle, and Christophe Dessimoz. Inferring horizontal gene transfer. *PLoS computational biology*, 11(5): e1004095, 2015.

[101] Jeffrey G Lawrence and Howard Ochman. Reconciling the many faces of lateral gene transfer. *TRENDS in Microbiology*, 10(1):1–4, 2002.

[102] Vincent Daubin, Emmanuelle Lerat, and Guy Perrière. The source of laterally transferred genes in bacterial genomes. *Genome biology*, 4(9): 1–12, 2003.

[103] Falk Hildebrand, Axel Meyer, and Adam Eyre-Walker. Evidence of selection upon genomic gc-content in bacteria. *PLoS genetics*, 6(9):e1001107, 2010.

[104] Jeffrey G Lawrence and Howard Ochman. Molecular archaeology of the escherichia coli genome. *Proceedings of the National Academy of Sciences*, 95(16):9413–9417, 1998.

[105] Diego Cortez, Patrick Forterre, and Simonetta Gribaldo. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and orfans in archaeal and bacterial genomes. *Genome biology*, 10(6):1–13, 2009.

[106] Shannon M Soucy, Jinling Huang, and Johann Peter Gogarten. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8): 472–482, 2015.

[107] Edward Susko. Tests for two trees using likelihood methods. *Molecular Biology and Evolution*, 31(4):1029–1039, 2014.

[108] David Schaller, Manuel Lafond, Peter F Stadler, Nicolas Wieseke, and Marc Hellmuth. Indirect identification of horizontal gene transfer. *Journal of Mathematical Biology*, 83(1):1–73, 2021.

[109] Christophe Dessimoz, Daniel Margadant, and Gaston H Gonnet. Dlight–lateral gene transfer detection using pairwise evolutionary distances in a statistical framework. In *Annual International Conference on Research in Computational Molecular Biology*, pages 315–330. Springer, 2008.

[110] Kenneth Katz, Oleg Shutov, Richard Lapoint, Michael Kimelman, J Rodney Brister, and Christopher O'Sullivan. The sequence read archive: a decade more of explosive growth. *Nucleic acids research*, 50(D1):D387–D390, 2022.

[111] Rufeng Li, Lixin Li, Yungang Xu, and Juan Yang. Machine learning meets omics: applications and perspectives. *Briefings in Bioinformatics*, 23(1): bbab460, 2022.

[112] Francesco Asnicar, Andrew Maltez Thomas, Andrea Passerini, Levi Waldron, and Nicola Segata. Machine learning for microbiologists. *Nature Reviews Microbiology*, 22(4):191–205, 2024.

[113] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[114] Subhash Nerella, Sabyasachi Bandyopadhyay, Jiaqing Zhang, Miguel Contreras, Scott Siegel, Aysegul Bumin, Brandon Silva, Jessica Sena, Benjamin Shickel, Azra Bihorac, et al. Transformers and large language models in healthcare: A review. *Artificial intelligence in medicine*, page 102900, 2024.

[115] Batta Mahesh et al. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, 9(1):381–386, 2020.

[116] Gireen Naidu, Tranos Zuva, and Elias Mmbongeni Sibanda. A review of evaluation metrics in machine learning algorithms. In *Computer science on-line conference*, pages 15–25. Springer, 2023.

[117] Davide Chicco, Matthijs J Warrens, and Giuseppe Jurman. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peerj computer science*, 7: e623, 2021.

[118] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction*, pages 109–139. Springer, 2022.

[119] Pedro Larranaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Inaki Inza, José A Lozano, Rubén Armananzas, Guzmán Santafé, Aritz Pérez, et al. Machine learning in bioinformatics. *Briefings in bioinformatics*, 7(1):86–112, 2006.

[120] K Aditya Shastry and HA Sanjay. Machine learning for bioinformatics. *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*, pages 25–39, 2020.

[121] Ewen Callaway. What's next for the ai protein-folding revolution. *Nature*, 604(7905):234–238, 2022.

[122] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[123] Luciano A Abriata. The nobel prize in chemistry: past, present, and future of ai in biology. *Communications Biology*, 7(1):1409, 2024.

[124] Yanjun Qi. Random forest for bioinformatics. *Ensemble machine learning: Methods and applications*, pages 307–323, 2012.

[125] Robin Genuer, Jean-Michel Poggi, Robin Genuer, and Jean-Michel Poggi. *Random forests*. Springer, 2020.

[126] Vinícius G Costa and Carlos E Pedreira. Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 56(5):4765–4800, 2023.

[127] Michele Fratello, Roberto Tagliaferri, et al. Decision trees and random forests. *Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics*, 1(S3):374, 2018.

[128] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[129] Adele Cutler, D Richard Cutler, and John R Stevens. Random forests. *Ensemble machine learning: Methods and applications*, pages 157–175, 2012.

[130] Philipp Probst, Marvin N Wright, and Anne-Laure Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery*, 9(3):e1301, 2019.

[131] Markus Loecher. Unbiased variable importance for random forests. *Communications in Statistics-Theory and Methods*, 51(5):1413–1425, 2022.

[132] Stefano Nembrini, Inke R König, and Marvin N Wright. The revival of the gini importance? *Bioinformatics*, 34(21):3711–3718, 2018.

[133] Kellie J Archer and Ryan V Kimes. Empirical characterization of random forest variable importance measures. *Computational statistics & data analysis*, 52(4):2249–2260, 2008.

[134] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8:1–21, 2007.

[135] Marco Angelini, Graziano Blasilli, Simone Lenti, and Giuseppe Santucci. A visual analytics conceptual framework for explorable and steerable partial dependence analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

[136] Brandon M Greenwell. pdp: An r package for constructing partial dependence plots. 2017.

[137] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[138] Maria-Elisabeth Böhm, Mohammad Razavi, Nachiket P Marathe, Carl-Fredrik Flach, and D G Joakim Larsson. Discovery of a novel integron-borne aminoglycoside resistance gene present in clinical pathogens by screening environmental bacterial communities. *Microbiome*, 8(1):1–11, 2020.

[139] Roland Leclercq, Rafael Cantón, Derek FJ Brown, Christian G Giske, Peter Heisig, Alasdair P MacGowan, Johan W Mouton, Patrice Nordmann, Arne C Rodloff, Gian Maria Rossolini, et al. Eucast expert rules in antimicrobial susceptibility testing. *Clinical microbiology and infection*, 19 (2):141–160, 2013.

[140] David Lund, Nicolas Kieffer, Marcos Parras-Moltó, Stefan Ebmeyer, Fanny Berglund, Anna Johnning, D G Joakim Larsson, and Erik Kristiansson. Large-scale characterization of the macrolide resistome reveals high diversity and several new pathogen-associated genes. *Microbial Genomics*, 8(1):000770, 2022.

[141] Michael N Alekshun and Stuart B Levy. Molecular mechanisms of antibacterial multidrug resistance. *Cell*, 128(6):1037–1050, 2007.

[142] Yongfei Hu, Xi Yang, Jing Li, Na Lv, Fei Liu, Jun Wu, Ivan YC Lin, Na Wu, Bart C Weimer, George F Gao, et al. The bacterial mobile resistome transfer network connecting the animal and human microbiomes. *Applied and environmental microbiology*, 82(22):6672–6681, 2016.